UNED
Escuela Internacional de Doctorado
EIDUNED

TESIS DOCTORAL

AÑO 2018

A Methodological Approach based on Machine Learning to Generate a Multimodal User's Affective State Model in Adaptive Educational Systems

Sergio Salmerón Majadas
Ingeniero Superior en Informática

PROGRAMA DE DOCTORADO EN SISTEMAS INTELIGENTES

Dr. Jesús González Boticario
Dra. Olga Cristina Santos Martín

Watch this, Lise.

You can actually pinpoint the second when his heart rips in half

Bart Simpson

A mis padres,
por hacer que todo pueda ir como siempre

Acknowledgments

I would like to thank / Me gustaría agradecer:

- *A todos aquellos honrados ciudadanos españoles que mediante el pago de sus impuestos han hecho posible la financiación de este trabajo.*
- *Mis directores de tesis por darme la oportunidad de llevar a cabo este doctorado.*
- *A mi familia (madre, padre, hermano, hermana, cuñada, tíos, abuelos y primas), por serlo, en todas las dimensiones posibles.*
- *A mis (ex)compañeros del grupo aDeNu, de departamento y de reuniones (Mar, Emmanuelle, Alejandro, Raúl, Raúl, Pilar, Emmanuelle, Elena, Mariano, Estela, Laura, Miguel, Felisa, Lourdes, etc.) por el soporte, formación y motivación dado a lo largo de estos años.*
- *A (entre otros) amigos Alberto y Marina, Aleix y Ainoa, Alexis, Guillermo y Olivia, Iván, Juanjo, Sergio, Clara, Cris, Sandra, Adrián, Lucia, etc.*
- *A los amigos que no he incluido en el punto anterior.*
- *Ryan S. Baker and Cristina Conati (And their groups: Mia, Dereck, Miggy, Jaclyn, Sweet, Shimin, Elle, Yang, Sebastien, Mike, Kailang, Giuseppe, etc.) for hosting me as a Visiting Researcher. Sharing time with all of you was such an enriching experience!*
- *The Krolik family for being my American family, dziękuję.*
- *The thesis defense committee members, international reviewers, my doctoral consortia advisors and all the attendees of my talks and poster presentations for the feedback provided.*
- *You for reading this.*

Abstract

The field of affective computing has been object of research over the last three decades. It focuses on the way electronic devices might interact with the emotional dimension of the device's users, detecting the affective state of the users in order to model it and use it with different purposes. In addition, there is strong evidence that emotions influence the learning process, which raises a huge number of potential applications of this affective computing field to help learners following a computer-based learning approach.

There is a growing number of research works focused on performing emotion detection by means of machine learning techniques. There is a wide variety of data sources proposed in literature in order to detect the affective state of users, from the commonly used physiological signals indicators to interaction devices such as keyboard or mouse. Regardless the data sources used, many works follow a similar approach in the affective state detection: i) collect data, ii) generate an affective labeling for the data and iii) use machine learning techniques to generate a prediction model. Despite of those points there is a clear lack of methodological comparison analysis in the literature, as most of the works propose an approach to perform that affective detection, but does not evaluate how each one of the methodological decision taken impacts on the results obtained.

In this Ph.D. Thesis, a research plan has been set in order to explore how to perform affective state detection using machine learning techniques (following a multimodal approach) and evaluate some of the different methodological issues faced in the design of that detection. For that, three different research stages were proposed: i) the first stage aims to perform an exploratory analysis on all the different methodological issues to research in the field of affective state detection from a multimodal approach in order to develop an initial experimental infrastructure; ii) a transition stage aiming to settle a reference context in order to drive the experimental approach followed in our first experiments to a more realistic scenario is carried out and iii)a final stage where the proposed methodological approach is adapted and evaluated in a real-world learning scenario, evaluating new methodological variables related to the kind of experimental approach followed (an inter-subject real-world learning scenario based experiment). During the experiments carried out, three different methodological dimensions where identified (i.e. characterizing and labeling affective state, data processing and experimental approach) and several methodological variables included in them were evaluated: the data sources to be used, different aspects from the affective data labeling performed to train the supervised learning algorithms used (from the labeler to the way to discretize the dimensional values collected), the data mining algorithms used, some preprocessing techniques used prior to the data mining algorithm model generation, etc. In addition, inspired by practice in affective computing where physiological sensors are used, a way to normalize interaction data according to each individual interaction skills has been proposed.

This work aims ultimately to define a methodology (named AMO-ML after Affective MOdeling based on Machine Learning) to perform affective state detection using machine learning techniques from a combination of different data sources. Additionally different methodological issues faced in the affective computing field are analyzed in three experiments. Also, the introduction of the interaction normalization approach seems to provide good results when predicting the affective valence (one of the dimensions of the affective state to evaluate) of the participants.

Resumen

El campo de la computación afectiva ha sido objeto de investigación durante las últimas tres décadas. Dicho campo se centra en cómo los dispositivos electrónicos pueden interactuar con la dimensión emocional del usuario, detectando el estado afectivo del usuario para modelarlo y que pueda ser utilizado de diversas formas. Además, existen estudios que establecen la influencia que las emociones pueden tener sobre el proceso de aprendizaje, lo que plantea un gran número de posibles aplicaciones que el campo de la computación afectiva puede tener para ayudar a estudiantes de plataformas de aprendizaje por ordenador.

Existe un creciente número de trabajos que se centran en la detección de emociones mediante el uso de técnicas de aprendizaje automático (o machine learning). Existe además una amplia variedad de fuentes de datos utilizadas en la literatura para detectar el estado afectivo de los usuarios, desde las comúnmente utilizadas señales fisiológicas a dispositivos de interacción como pueden ser el teclado o el ratón. Independientemente de la fuente de datos que utilicen, muchos trabajos siguen un enfoque similar en la detección de estados afectivos: i) recoger datos, ii) generar un etiquetado afectivo para esos datos y iii) utilizar técnicas de aprendizaje automático para generar un modelo predictivo. A pesar de esos puntos comunes, hay una clara falta de análisis comparativo en las metodologías de la literatura relacionada, ya que la mayor parte de los trabajos proponen un enfoque para llevar a cabo dicha detección del estado afectivo, pero no se evalúa el impacto de cada una de las decisiones metodológicas tomadas en los resultados obtenidos.

En esta tesis doctoral se ha establecido un plan de investigación para explorar como llevar a cabo la detección de estados afectivos mediante el uso de técnicas de aprendizaje automático (a partir de un enfoque multimodal) y evaluar algunos de los puntos metodológicos afrontados en el diseño de dicha detección. Para ello, se han propuesto tres fases en la investigación: i) en la primera fase se lleva a cabo un análisis exploratorio sobre los distintos puntos metodológicos en la investigación dentro del campo de la detección del estado afectivo desde un punto de vista multimodal para poder llevar a cabo una infraestructura experimental inicial; ii) una fase de transición para establecer un contexto de referencia para guiar el enfoque experimental de los primeros experimentos hacia un escenario más realista y iii) una fase final en la que el enfoque metodológico propuesto es adaptado y evaluado en un escenario realista de aprendizaje, evaluando las nuevas variables metodológicas relacionadas con el enfoque propuesto (un experimento inter-sujeto basado en el entorno de aprendizaje realista). Durante los experimentos llevados a cabo, se han identificado tres dimensiones metodológicas (i.e. caracterización y etiquetado de los estados afectivos, procesado de datos y enfoque experimental) y diversas variables metodológicas incluidas en dichas dimensiones han sido evaluadas: las fuentes de datos a usar, diversos aspectos del etiquetado afectivo de los datos para entrenar los algoritmos de aprendizaje supervisado utilizados (desde el etiquetador hasta la forma en la que se discretizan los valores dimensionales recogidos), los algoritmos de minería de datos utilizados, algunas técnicas de preprocesado aplicadas antes de la generación de los modelos de minería de datos, etc. Además, inspirada en una práctica dentro del campo de affective computing con señales fisiológicas, se propone una forma de normalizar los datos de interacción en base a las habilidades de interacción de cada individuo.

Este trabajo pretende, fundamentalmente, definir una metodología (llamada AMO-ML, siglas en inglés de MOdelado Affectivo basado en Aprendizaje Automático) para llevar a cabo predicción de estados afectivos mediante técnicas de aprendizaje automático sobre una combinación de diversas fuentes de datos. También se analizan diferentes aspectos metodológicos encontrados en el campo de la computación afectiva en tres experimentos. Además, la introducción del enfoque de normalización ofrece buenos resultados en la predicción de la valencia (una de las dimensiones a evaluar de los estados afectivos) de los participantes.

Preface

All of the work presented henceforth was conducted in the Laboratory for Affective Computing and Inclusive Interaction of the aDeNu Research Group, a multidisciplinary group from the Artificial Intelligence department in the School of Computer Engineering at Universidad Nacional de Educación a Distancia (UNED), specialized in the development of adaptive interfaces via Internet, based on user modeling through a combination of machine learning techniques. With this background, the work presented is framed in the projects: i) MAMIPEC (Multimodal approaches for Affective Modeling in Inclusive Personalized Educational scenarios in intelligent Contexts - TIN2011-29221-C03-01) [209], funded by the Spanish Ministry of Science and Innovation under the Subprogramme for non-oriented fundamental research projects and its successor, ii) BIG-AFF (Fusing multimodal Big Data to provide low-intrusive AFFective and cognitive support in learning contexts - TIN2014-59641-C2-2-P) [34], funded by the Spanish Ministry of Economy and Competitiveness under the RD projects call in 2014. In these projects, a number of key issues in the fields of affective computing, context awareness and ambient intelligence have being addressed, studying their application in adaptive and inclusive educational contexts [35]. MAMIPEC project included a FPI grant (BES-2012-054522), which is funding the research work presented in this document. That FPI grant (from the Ministry of Economy and Competitiveness) has provided 4 years of funding for this research, as well as funding for the MSc and Ph.D. fees and the two research visits performed during this research.

The work here presented started in the beginning of the MAMIPEC project, and it is the continuation of a Master's Degree Thesis (done in the frame of the UNED's University Master's Degree in Advanced Artificial Intelligence: Principles, Methods And Applications), which received a special mention from the eMadrid Consortium. During the elapsed time of the projects, the research work of this Thesis has been presented in a JCR Q1 journals [193] (IEEE Access) and another article is to be resubmitted to JCR indexed journal (Transactions on Intelligent Systems and Technology) as well as in international research conferences related with user modeling, data mining, artificial intelligence and learning technologies (EDM 2013 [201], AIED 2013 [212], HCII 2013 [211], EDM 2014 [197], KES 2014 [198], AIED 2015 [192]). In addition the research approach was accepted and discussed in several Doctoral Consortia (UMAP 2013 [196], EDM 2013 [194], CAEPIA 2013 [195], UMAP 2014 [191], AIED 2015 [200]) where the work presented was found of interest and of relevance, and valuable feedback was provided by relevant researchers from the field. All the feedback received has been used to refine the proposal of this work. In this sense, it has to be highlighted the mentors assigned in UMAP Doctoral Consortia: Mária Bieliková[1] and Marko Tkalčič[2]. Thanks to the funding of the project and in some cases, external funding[3], I was able to attend all the aforementioned conferences in order to present the research work carried out and take advantage of the opportunity of sharing my work and getting to know in person the state of the art in the field as well as the big names in it. All these conferences are indexed in CORE ranking (A o B), except CAEPIA (which in turn is the main Spanish conference in the field and very well positioned in specialized international rankings) and HCII. The feedback received in the doctoral consortia held by the PhD program where this work is presented has also been a huge help in order to guide the direction of this work, as well as the yearly reports generated by the doctoral commission. A yearly doctoral consortium was also programmed within the Intelligent Systems Ph.D. program, so this work was also presented in Jornadas de Doctorado 2015[4] (with the advice of Dr. Milos Kravcic from Aachen University, Dr. German Rigau from UPV/EHU and Dr. Maria Süveges from University of Geneva) and Jornadas de Doctorado 2016 (with the advice of Miriam Fernández from the Knowledge Media Institute at the Open University and Dr. Roberto Moriyón from UAM).

During this PhD program, two different research visits (both funded by the Spanish Ministry of Economy and Competitiveness trough the FPI program) were made: first one in 2015 visiting Dr. Cristina Conati, head of Intelligent User Interfaces Research Group at University of British Columbia in Vancouver, Canada (funding ID: EEBB-I-15-10414). The second visit was made in 2016 under the supervision of Dr. Ryan Baker, head of the Educational Data Mining Lab at Columbia University in New York City, USA (funding ID: EEBB-I-16-11857).

Moreover, this research work has been reported as part of other publications that disseminate findings of the MAMIPEC project, as follows: i) a journal paper in a JCR indexed publication Q2 [204], ii) research conferences, such as ICALT 2014 [214] and the workshop PALE [10,11], iii) teaching innovation conferences, such as Jornadas de Innovación Educativa of the Universidad of Valencia [8] and Jornadas de Redes de Investigación en Innovación Docente of UNED [35,109].

With Jesús González Boticario and Olga C. Santos as doctoral advisors, the author was the main contributor, designer and implementer of all the works carried out in this Doctoral research, including the AMO-ML methodology designed, which is reflected in this manuscript and resulting works.

[1] ***Mária Bieliková*** is a full professor at the Slovak University of Technology in Bratislava with a huge background on e-learning, member of IEE and its Computer Society, ACM or ISWE among others.

[2] ***Marko Tkalčič*** is a post-doctoral researcher at the Department of Computational Perception in Johannes Kepler University Linz (Austria), with a research line centered on affective computing, organizer of the workshop series on "Emotions and Personality in Personalized Services" (EMPIRE) run since 2013 at UMAP conference (CORE B).

[3] User Modelling inc.

[4] Video available at: https://canal.uned.es/video/5a6f6cf4b1111f26508b459f

A list of all the resultant publications of this work as well as collaborations in other publications related to the work here presented can be found in Appendix I (section 13.1).

MAMIPEC Project

MAMIPEC (Multimodal approaches for Affective Modeling in Inclusive Personalized Educational scenarios in intelligent Contexts) project was a research project funded by the Spanish Ministry of Science and Innovation (with ID TIN2011-29221-C03-01). The project was carried out by the aDeNu research group in collaboration with a research group in University of Valencia. The project was initially scheduled from 01/01/2012 until 12/31/2014, but it was extended until 06/30/2016.

In this project, the goal was to address a number of key issues in the fields of affective computing, context awareness and ambient intelligence. In particular, to study their application in adaptive and inclusive educational contexts. This implies the introduction of new and more complex modeling needs which have not been considered in most previous learning research, along with a broad investigation of related topics, including a) affective interaction by means of natural implicit and explicit interfaces; b) information processing to assess the user’s state by using multi-modal approaches; c) the inclusion of affective information in the user model; d) environment/context modeling; e) the provision of adaptive behavior; and f) the integration of ambient intelligence in learning.

Indeed, affective computing, context awareness and ambient intelligence applications are not limited to learning. Although one major concern of this proposal is to increase understanding on effective methods for exploiting these concepts to benefit learning, the results of this research may be extended to other relevant application areas of the human-computer interaction field such as supporting the independence of people with special needs or reducing the digital divide facilitating the integration of minorities (migrants, reduced literacy people, etc.).

All the work here presented has been done within the frame of the MAMIPEC project, as the funding of this research is based on a FPI grant attached to this project. That FPI included the funding for 4 years.

Another important input from this project to the research here presented is the inclusion of University of Valencia, which has provided a valuable help in the works performed. This collaboration also has driven some aspects of this research, such as the inclusion of some tools developed by them. One of these tools is a tool for recording Kinect facial data (mentioned in section 4.5.4), allowing to export facial features recognized by the Kinect device to csv files. Although the previous tool was used in the experiments described in sections 4 and 5, one of the most important contributions of the research carried out by University of Valencia to this work is the use of an Intelligent Tutoring System (described in section 5.2.4) developed by them. That ITS, centered on algebra problem solving was used as the task to be solved by the participants in the experiments described in section 5.

BIG-AFF Project

BIG-AFF (Fusing multimodal Big Data to provide low-intrusive AFFective and cognitive support in learning contexts) project was a research project funded by the Spanish Ministry of Science and Innovation (with ID TIN2014-59641-C2-2-P). The project was the continuation of the MAMIPEC project, and was carried out, again, by the aDeNu research group in collaboration with a research group in University of Valencia. The project was initially scheduled from 01/01/2015 until 12/31/2018.

This project was born as the continuation of the MAMIPEC project, aiming to "provide learners with a personalised support that enriches their learning process and experience by using low intrusive (and low cost) devices to capture affective multimodal data that include cognitive, behavioral and physiological information". Part of the work here presented has been reported in this project.

Additionally, during this project some hardware developments have been carried out. Some of these hardware developments have been used in some stage of this work (described in section 6.5.1.c).

Index

List of Tables

List of Figures

1. Introduction

1.1.Motivation

Although is in these years when we can see a boom in distance learning courses (especially nowadays with the so called Massive Open Online Courses - MOOCs), these courses have been present for centuries. It was in the eighteenth century when Caleb Phillips published an advertisement in the Boston Gazette [36] in order that "any persons in the country desirous to learn this Art, may, by having the several lessons sent weekly to them, be as perfectly instructed as those that live in Boston" could learn his shorthand method. This methodology was reproduced by Isaac Pitman's in mid nineteenth century, when he created correspondence delivered courses to teach his shorthand system and, some years later, the Phonographic Correspondence Society (the beginning of the Sir Isaac Pitman Colleges across the country) [36]. In Germany Charles Toussaint and Gustav Langenscheidt developed some self-learning methods to teach languages ("Teaching letters for learning the French language" or "English lessons letters"), having their own publishing group with a printing press [111]. In the following years, the creation of radio and television stations was also used with educational purposes, appearing in the first 30s the first television shows scheduled by an university in Iowa [156].

In Spain, the clearest example of distance learning is the Universidad Nacional de Educación a Distancia (UNED). When created (i.e., in the 70s) teaching was provided by sending by postal mail the materials to the learners and getting their responses back in the same medium. Questions could be solved by phone. During the 80s, this distance learning approach was reinforced by the use of radio and television to broadcast some materials. It was in the 90s when UNED started using multimedia systems not only for creating contents but also for its dissemination, becoming a pioneer in Spain in the use of technology for distance education. Nowadays UNED relies on a distance education approach able to handle more than 260.000 learners being also the university with more students with accessibility requirements in Spain [256].

It was with the appearance of computers and the dissemination of the Internet, when distance learning has become most popular and can be more adaptive and personalized, as some of the problems in previous distance learning approaches allow to break, in a more efficient manner, the time and space barriers.

It is in the last 1980s and in the 1990s, with the popularization of computers [160] for personal or home use, when computer-based courses became prominent. Initially, they offered almost no interaction, but this is something that has gradually changed [98].

This was the beginning of the so called e-learning. In the last 90s, with the spread of the Internet, web courses emerged becoming in the 2000s a common tool used by universities all over the world [48]. In the current decade, the creation of MOOCs is democratizing learning, allowing anyone with Internet access to enroll in thousands of courses offered by many different and prestigious academic institutions [250].

The techniques involved in distance education have evolved during the last three centuries, but the appearance of e-learning is the first opportunity in the distance education field to offer a real interactive, collaborative and adaptive experience [98], trying, this way, to emulate a traditional classroom approach and even going beyond, looking for a fine-grained personalization even hard to provide in a traditional classroom [66].

According to Bernard Luskin, one of the e-learning pioneers, the “e” in e-learning does not stand only for “electronic” as it should mean also "exciting, energetic, enthusiastic, emotional, extended, excellent, and educational" [108]. These are some of the directions to advance in the e-learning field, and the “emotional” term is concretely the one that motivates this work. Furthermore, is in making e-learning an adaptive experience where Psychology plays an important role as this adaptation should be driven by the learning process, where a lot of psychological aspects take part on it. One of those aspects is the emotional feature of learners, which play a key role on learning process. The emotional aspects of learning are a factor that has been increasingly taken into account in learning process [68]. Despite of that, teachers are usually not taught how to address affective issues when dealing with students [38], and this gap is even bigger in distance learning courses [38,143].

Nevertheless, the emotional dimension of a user while interacting with a device is something that has been widely explored in the last years, studying its application in many different fields, from videogames [87,89] to online marketing [146]. Due to the existing relations between emotions and cognitive processes in learning [168,182], e-learning is a field that is can also advance in the direction of the affective states detection to take advantage of it [60,165]..

It was in the mid-90s when Rosalind Piccard used the term “affective computing” for the ability of computers to recognize and express affect [174]. Many works have been carried out in both of these directions. On the one hand, detecting the affective state is a complex process that can be addressed from different perspectives, as commented next and further detailed in the state of the art reported in Section 2. On the other hand, providing appropriate affective support has many open issues, some of them regarding distance education have been discussed elsewhere [143]. Providing an appropriate affective support requires that the emotions on the learner are accurately detected. For this reason, this work focuses on how to improve affect recognition in the educational domain.

By means of detecting, modeling and processing the affective state of a human a system can try to interact with psychological aspects related to emotions, such as motivation [59]. Many different ways can be considered to make a system aware of the emotions experienced by its user, going from asking the user about his or her feelings using questionnaires [169] to automatic emotion recognition systems based on the analysis of some behaviors or signals from the user (such as facial expressions [204], physiological measures [89] or behavioral changes in interaction while interacting with devices [62,78,84]).

During last years, technological advantages have made possible the development of new devices that facilitate the measurement of some physiological variables out of a clinical environment [51,69,184,206], and even in wearable devices, allowing non-intrusive ways of collecting physiological signals [224]. One of the most common approaches nowadays in emotion detection is the use of physiological sensors, which usually involves capturing huge amounts of data. Due to the volume of the data generated by those devices, the machine learning techniques capable to extract information from that data has been identified as a reasonable approach to face the processing of the data collected [251]. This is due to both the volume and the complexity of the data to analyze, due the importance of analyzing as many emotional events as possible, by means of creating a dataset as rich as possible to be used with data mining techniques. Is this reason, the need of an Intelligent System capable to do all this why this research has been framed in the Intelligent Systems Doctoral Program. For the implementation of that intelligent system, the methodology named AMO-ML (Affective MOdeling based on Machine Learning) has been proposed.

To summarize, this work is motivated by the combination of several factors: i) the role that emotions play in education [119]; ii) the capabilities that adaptation and personalization in e-learning scenarios can provide to the learner [163]; iii) the growing research in affective computing [56,167] and advances in emotion detection during the last years [65]; and iv) the capabilities of machine learning and data mining when handling data with affective purposes [180].

1.2.Problem Definition

As aforementioned, emotional aspects of learning are a factor that has been increasingly taken into account in learning process [68]. These aspects are to be used by e-learning scenarios in order to improve the learning experience of the users by means of providing affective feedback. In order to provide that feedback, it is needed to detect the current affective state of the learner, and it is in that emotional detection process where remains the main problems to be faced in this work. The goal of the work here presented is to provide a methodology (named AMO-ML after Affective MOdeling based on Machine Learning) to detect and model emotions from data collected following a multimodal approach by means of machine learning techniques. To develop that methodology, all the methodological issues found in the design and development of a system capable to perform that detection have to be identified.

Many works have been carried out during the last years aiming to perform emotional states detection, following a wide variety of approaches [65,180] and only a subset of these works are focused on the educational field [205]. When evaluating the different works in literature, it can be seen that most works aim to propose an approach in emotion detection. This results in a wide variety of approaches, each one aiming to detect emotions, but a lack of evaluation on how the different methodological issues faced in these works may have an impact on the results obtained.

With that wide variety of studies in mind, the main problem to be addressed is how to perform affect detection in educational environments by combining information gathered from several input sources with a multimodal data mining approach, evaluating the different methodological questions involved in performing that detection. This requires an initial study and identification of all the related issues in the affective state detection in order to structure the following steps to be taken.

Once these issues have been identified, an initial methodology has to be developed taking into account the related literature. After that, different analyses of a series of research issues found during the construction of that approach are going to be faced. The first one is the data sources to be used. The different approaches found in literature usually introduce a single or a set of data sources to use, but they do no traditionally evaluate the prediction results provided by every possible combination of data sources, evaluating which data source combination might be the best for affective state detection. Regarding the data sources used, some aspects such as intrusiveness have to be taken into account, aiming to frame this research in a as realistic as possible context. The second one is how to handle the information coming from those data sources, for instance, an interaction data normalization which aims to propose a way to normalize data collected from mouse and keyboard using as reference point values collected from each subject. This might be helpful to use those data sources in inter-subject (i.e. merging data from several subjects) experiments where the diversity of typing or mouse movement skills might have a negative impact on the models generated from raw data. The main data sources to be evaluated in this work include: interaction devices (i.e. keyboard and mouse) and physiological signals. Other data sources to be used in this work include sentiment analysis and task performance related features.

In addition to those problems to be addressed, other research issues are going to be addressed. These issues are based on the evaluation of the impact some methodological issues might have on the results of the affective state detection. These issues include the impact different preprocessing techniques (i.e. class balancing techniques and dimensionality reduction techniques) might have on the models generated. This data preprocessing process is an issue that is not commonly described in a detailed way in related works, and is going to be evaluated in this work. Another methodological issue to be faced is the emotional labeling to be performed, evaluating different approaches (from the labeler to the emotional representation used). Nevertheless, some points commonly addressed are also going to be researched, such as the impact the data mining algorithm used might have on the prediction results.

1.3. Proposed approach

The global framework of this work relies on the affective computing area. In this area, the work here presented aims to research in the development of emotion detection systems based on machine learning as well as the impact of the methodology followed on the design and use of those systems. The different methodological issues to be evaluated in this work show different abstraction levels, evaluating the coarse-grained issues in the first stage of this research during the first stage and the finer-grained methodological issues in the final stage. The context of application of the research here presented is the educational, and will drive the design of the experimentation materials to be used. Regarding the tools to perform the prediction, machine learning algorithms (more concretely, supervised learning techniques) are to be used to process all the data collected during the experiments.

The approach here proposed aims to deal with several global purposes that remain underlying in the hypothesis and objectives of this work (described in section 1.4). These purposes include:

- Research in the field of affective state detection on the fly in real world learning scenarios, using machine learning techniques.
- Study potential benefits of using several data sources in order to provide the best indicators from those data sources for each context, taking advantage of that data source variety.
- Identify and propose potential solutions to those methodological aspects that arise in approaches based on the combination of different data sources with affective state detection purposes.
- Define an effective solution in order to deal with the methodological variables identified (studying how those variables should be used in the process proposed).
- Study different affective labeling approaches aiming to maximize their effectiveness in different use cases. Points such as data sources intrusiveness and impact on the task being performed should be taken into account.
- Develop tools that support the affective state detection dealing with the modeling problems involved.
- Take into account the adaptability of the proposed approach into multiple scenarios. In this sense the most common interaction data sources used in e-learning (i.e. keyboard and mouse) as well as a combination of other different data sources commonly identified in related works.

Also, the proposed approach is to be driven by a series of questions that will be the base for the research hypotheses (to be introduced in section 1.4) to face in this work. The research questions faced in this work include:

- **Q0**: Can machine learning techniques be used in order to detect learner’s affective states in realistic learning scenarios from data collected from different data sources?

- **Q1**: Can the combination of different data sources in educational scenarios help to improve the affective state detection compared to single-data source approaches?
- **Q2**: Which are the methodological aspects involved in the use and combination of different data sources with affective state detection purposes?
- **Q3**: Which affective state labeling strategies are more effective in real-world educational scenarios without penalizing aspects such as the intrusiveness of the approach proposed?
- **Q4**: Can the use of reference interaction patterns (collected in a non-affective task) reflecting each individual personal interaction behavior help to improve the affective state detection in real-world learning scenarios?
- **Q5**: When using reference interaction patterns, can the regular update of those reference patterns help to improve the affective state detection in contrast to using a reference interaction pattern at the beginning of the interaction?
- **Q6**: To what extend the way the multimodal data collected in a real-world learning scenario is handled prior to the model generation with affective state detection purposes can have an impact on the prediction results obtained from that model?

To address the proposed questions a series of explicit objectives have been defined in this research (to be introduced in section 1.4), with some more generic questions, such as Q0, which is addressed all over the research here introduced.

In order to face all the purposes and questions of this research, the approach proposed in this work has been designed in two main stages, each one with a clear research hypothesis and objectives (that address the different research questions introduced). Additionally, a small transition stage has been included between those two main stages. The idea is to follow an incremental approach, building the second stage from the results of the first one.

- The first stage of this research aims to perform a deep exploratory analysis of the affective state detection in educational scenarios field. This analysis aims to draw an initial methodological approach to perform emotion detection, as this is the main tool to be used for evaluating the hypotheses proposed. To get there, a configuration has to be designed according to the systems proposed in the literature, proposed and evaluated in a first experiment performed in controlled conditions. In this stage, in addition to the evaluation of the basic settings and performance of the system developed, other methodological aspects identified in the initial research performed in this stage are to be evaluated, facing this way the research questions **Q1**, **Q2** and **Q3**.
- A small transition stage is going to be carried out after the initial stage has been finished. At the end of that first stage, a clear view of the field has been translated into an affective state detection methodology. With that initial approach defined, the definition of a real-world based scenario in order to apply the designed methodology has to be done. The goal of the inclusion of areal

world learning scenario is to provide ecological validity to the methodology here proposed. This step will be taken in a transition between first and second main stages of this research work. A small experiment will be carried out to identify aspects that should be taken into account in the application of our research in a classroom.

- The second and last stage of this research focuses on the methodological development of the approach defined in the stage 1 and its application on a real-world learning scenario (defined in the transition stage). The work in this stage starts with the approach defined in the first stage and the scenario proposed in the transition stage. The main goal of this stage is to improve the methodology created in the previous stage by researching some finer-grained methodological aspects found in the application of the methodology proposed in experimentations performed in real-world learning scenarios. The methodological variables evaluated at this point are related to the data processing, aiming to discuss the impact these different processing approaches might have on the results obtained. The issues addressed in this second stage include those involved in the research questions **Q4**, **Q5** and **Q6**.

1.4. Hypotheses and Objectives

The main objective of this research work is the development of a methodology to perform affective state detection by means of machine learning techniques in a multimodal approach based in learning scenarios. In that direction, with the research structure proposed in the previous section, the following hypotheses and objectives have been proposed:

Stage 1 Hypothesis (**H1**)

> *Supervised data mining techniques on multimodal data sources improve the accuracy when detecting affective features to enrich learner modeling in task-independent educational contexts in comparison with single data sources.*

Stage 1 Objectives (**O1**)

- *Evaluate different non-intrusive data sources to be used on emotion detection (**O1.1**).*
- *Evaluate different emotion labeling approaches to be used as dependent variables (**O1.2**).*

Stage 2 Hypothesis (**H2**)

> *In real-world learning scenario based inter-subject experiments, the use of a reference state to normalize each user interaction related data provides more robust models when detecting affective features to enrich learner modeling in educational contexts.*

Stage 2 Objectives (**O2**)

- *Evaluate the impact of user centered normalization in across subject experimental approaches (**O2.1**).*
- *Evaluate different preprocessing techniques on the data collected and their impact on the results (**O2.2**).*
- *Evaluate different data discretization approaches on affective numerical labels (**O2.3**).*

1.5. Research Methodology

In order to evaluate the hypothesis, an experimental-based methodological approach has been proposed. In particular, the following steps have been set for each one of the two iterations proposed:

1. Review the state of the art in order to have a clear view of the field of emotion detection in computing, detecting open issues, successful approaches and research lines to follow.
   a. In stage 1, the focus of this step is to be set on the design and development of an initial version of the methodology to follow during the experiments to detect the affective state of learners.
   b. In stage 2, the focus of this step is to be set on the application of the methodology and experimental design from previous stage in real-world learning scenarios, as well as on the data preparation steps followed by other related works.
2. Select data sources to be used, detecting, how the data should be extracted from them and how this data should be processed in order to get affective information.
   a. In stage 1, this step has to result in a series of data sources to be used (taking into account factors such as intrusiveness, price, availability, etc.) due to their potential value providing affective information.
   b. In stage 2, this step aims to evaluate the use of the data sources used and the methodology designed in the previous stage in a real-world scenario.
3. Propose a work plan to follow and an experimental design.
   a. In stage 1, create and design the infrastructure needed to perform the affective state detection. Design an experiment capable to evaluate the performance of the system as well as the impact different methodological aspects might have.
   b. In stage 2, review both the infrastructure and the experimental design in order to take the existing experimental design to a real-world scenario. From the experimental design, identify the changes required in order to evaluate the methodological variables to be analyzed in the experiments to be performed.

4. Carry out an experiment to get the data generated from the selected data sources.
    a. In stage 1, hold an experiment in lab conditions in order to evaluate the different issues from the experimental set up that might impact on the results obtained.
    b. In stage 2, transfer the approach of the previous experiment to a real world educational scenario to evaluate the application of the previously proposed approach to real world conditions.
5. Prepare the data to be used by supervised learning techniques and generate different models from that data.
    a. Generate a data analysis workflow capable to perform affective state detection from the data collected in the previous step.
    b. Modify the data analysis workflow in order to include some analysis steps as methodological variables.
6. Analyze the results from the different models generated in order to evaluate the affective detection capabilities of these models, evaluating the impact the different methodological aspects studied in each stage have had on the results.

In this way, a whole research cycle was carried out in each iteration in order to evaluate the potential offered by different data source combinations to provide affective information of a learner while interacting with an e-learning platform.

1.6. Expected outcomes

The proposed work aims to advance in some aspects from the traditional approach adopted when performing affective states detection. The main outcome of the work here presented is a methodology named AMO-ML, based on an evaluation of the impact different methodological aspects involved in the development of an affective state detection system from different data sources using machine learning techniques. This aims to provide a data-based informed motivation to adopt or discard different aspects in the design of an affective state detection system.

In the first stage of this work, the methodology is to be designed and some wide range methodological issues detected are going to be evaluated. The development of a first experiment and the description of all the steps carried out from the experimental design to the machine learning based model generation give form to a methodology developed from the prior exploration of the related literature. In addition, due to the wide variety of data sources proposed in literature, different approaches have been proposed for that experiment. Going from physiological sensors to postural behaviors, but taking into account other possibilities, as nowadays the ways to access to learning platforms are been enhanced in terms of a wide variety of interaction devices, different data sources from interaction behaviors can be used to get affective information from the learner. Usually, works seen in literature (see section 2) have focused on studying one or a concrete closed group of sources. A high detailed study of all the possibilities combining sources is to be done in order to offer a more accurate affective detection that can serve to provide affective adaptation regardless of the available data sources available for a learner in a given context. Another important methodological variable evaluated in the first stage is the comparison between different affective labelers, aiming to provide information that might help in the decision making process when designing the source of the labels to train an intelligent affective state detection system.

During the second stage, an evaluation of the applicability in a real-world learning scenario of the methodology designed in the first stage is going to be carried out. In addition, a series of methodological aspects identified are going to be evaluated. Some common techniques traditionally used in data processing are related to problems commonly found in the research of affective state detection systems (i.e. small datasets, unbalanced class attribute, etc.). Also, the evaluation of the use of a reference value for normalizing interaction data (collecting, for instance, those reference values as a baseline as it is traditionally done in physiological signal analysis) has to be proposed and evaluated. By mean of this, this work aims to provide a way to improve the interaction data performance when used with affective purposes. This outcome is supposed to be especially valuable in inter-subject experimental approaches, as its main goal is to get rid of the user skill variable when normalizing the data (using each user interaction baseline as reference) of users with very different interaction skill levels. Another outcome to evaluate is the applicability of the proposed approach to a real-world learning scenario, describing the infrastructure and methodology followed to perform the experimentation in a real classroom.

The outcomes already mentioned draw a global outcome which is a clear methodological approach, applied in a real-world educational scenario and based on the research of several variables. This work is, in summary, the AMO-ML methodology proposal itself, but with methodological question identified in different steps of the proposed approach. This identification of open issues and the evaluation of some of them, aims to help in the advance of the affective computing field.

In addition to the main outcomes related to the goals proposed, other outcomes are going to result from this work: i) the design and implementation of a affective state detection system here depicted, where all the methodological aspects not evaluated are also going to be indicated (with some of the developments carried out in this work to be uploaded to a repository by the author); ii) the design of the experiments carried out is also an outcome, describing from the emotion elicitators used to the different steps taken in the experimental procedure; iii) a review of the state of the art, focusing the attention to the methodology followed in the development of different intelligent emotion detection approaches, aiming to provide a summary of the different decisions taken in the literature in this field.

1.7.Document Structure

The research work carried out and reported in this Thesis is structured as follows:

In section 2, a review of the state of the art on extracting affective information from different data sources shows the different approaches studied in order to get an overview of the field of affective state detection before creating a proposal of a new approach. It also analyzes the different data labeling approaches that can be used to apply supervised machine learning methods.

In section 3, the approach followed in this research work is introduced, defining the different stages this work has been structured in.

In section 4, the first stage of this research (or exploratory stage) is described. In this stage, an initial exploratory study is carried out, introducing a multimodal data mining approach to perform affective state detection, looking for different methodological open issues in the creation of this approach and its application in a first experiment.

In section 5, a transition stage to define a reference scenario to adapt the approach introduced in the exploratory stage to its application in a real-world context is presented. This transition is done by carrying out an experiment in a classroom to identify the problems that this new context may induce in the approach proposed.

In section 6, the methodological development of the different iterations performed is carried out, evaluating some of the issues found during the previous stages in a real-world context experiment. Special attention to data processing related issues are evaluated in this section, as well as the proposal of an interaction baseline to normalize data collected from some data sources.

In section7, a summary of the results achieved in this research work is included.

In section8, discussion about some of the issues found during the development of the approach here presented is introduced.

Section 9 contains a list with all the contributions generated from the work here presented.

Section 10 includes the conclusions outline of this work.

In section 11, different ways to continue the research here described are outlined. As this work is framed in a newfangled field, and due to the lack of strong methodology in it, some proposals of future works are introduced in this section.

Section 12 includes all the bibliography used in this work

Last section (Section 13) includes some appendices with materials used in this work

2. State of the Art

In order to develop an affective state detection methodology and evaluate some of the aspects involved in its design, it is necessary to carry out a careful analysis of the works that have been already performed by previous researchers. As this research has already been limited with some clear methodological points, the literature to use as the base of this work has to be framed in the area delimited by those constraints.

A general search of related works in affective state detection was performed, selecting those works which met some of the following points: i) works that focus on affective states detection, ii) works that use data sources considered in this research work (i.e. psychophysiological data sources and computer interaction devices as keyboard and mouse), iii) works that describe the feature generation process from the data collected, iv) works that use machine learning techniques in order to perform prediction, v) works that describe the data collection process, vi) works that, when using supervised learning techniques, describe the labeling approach of the data collected, vii) works that have been carried out in an educational context.

It is important to have a clear view of the information to extract from the selected works. With the hypothesis and objectives introduced in section 1.4, the focus is going to be set in issues that will be addressed in the stage 1 in order to face the objectives set for that stage, such as the data sources to be used (section 2.1); in stage 2 to address the objectives for that stage, such as the data analysis technique used (section 2.2); or in both sections such as the labeling (section 2.3) or the emotional elicitation method used (section 2.4).

It should be noted that the each study analyzed here provides different levels of information, so this review can only describe each study based on the information provided. Nonetheless, in order to extract general themes across studies, Table 2 and Table 3 are included after our analysis to summarize the different methodological aspects and the features used across studies.

The following sections include the conclusions from the review of related literature: i) data sources used in related works (section 2.1), ii) different data analysis technique used in related works(section 2.2), iii) labeling approach followed and different methodological aspects involved in labeling (section 2.3), iv) emotional elicitation method used (section 2.4), v) compilations of related works (section 2.5) and vi) conclusions of the related works analyzed (section 2.6).

2.1. **Data sources**

In this section it is going to be presented a review of the most relevant works that carry out affective states detection using the data sources considered in this research work. In Table 2 it can be found a summary of the data sources used in the related works here analyzed. At the initial stage of this review of the state of the art, the data sources to be used were not clear, so, the choice of the data sources to be used in our work was a result from this research task. This section has been structured in the two different main data sources families to be used in our work: psychophysiological signals and interaction devices (i.e. keyboard and mouse).

2.1.1. Psychophysiological sensors

One of the most common approaches on emotion detection relies on the use of Psychophysiological sensors, which aim to measure mechanical, physical, bioelectrical, or biochemical changes that occur in the human body. Many related works aiming to perform affective state detection consider different physiological measures which have been studied from the psychological field to have a link with affective states, including: electrocardiography based measures (which is based on the activity of the heart) [231], facial expression analysis (studying the behavior of muscles in the face) [75], electrodermal activity (based on the variations of the electrical conductivity of the skin) [164], breathing [29],skin temperature [149], voice analysis [112], electromyography (analyzing the electrical changes in muscles) [106], electroencephalography (that monitors the electrical activity of the brain) [102], electrooculography (measuring the potential that exists between the front and the back of the human eye) or eye tracking (another less intrusive way to measure eye movements and other eye related features) [104], etc. There is wide research using these signals as it is known the effects of emotions on the autonomic nervous system (that regulates functions as heart rate, respiratory rate, facial muscles, etc.) [74,132,179]. Figure 1 shows how psychophysiological features are widely used in the field.

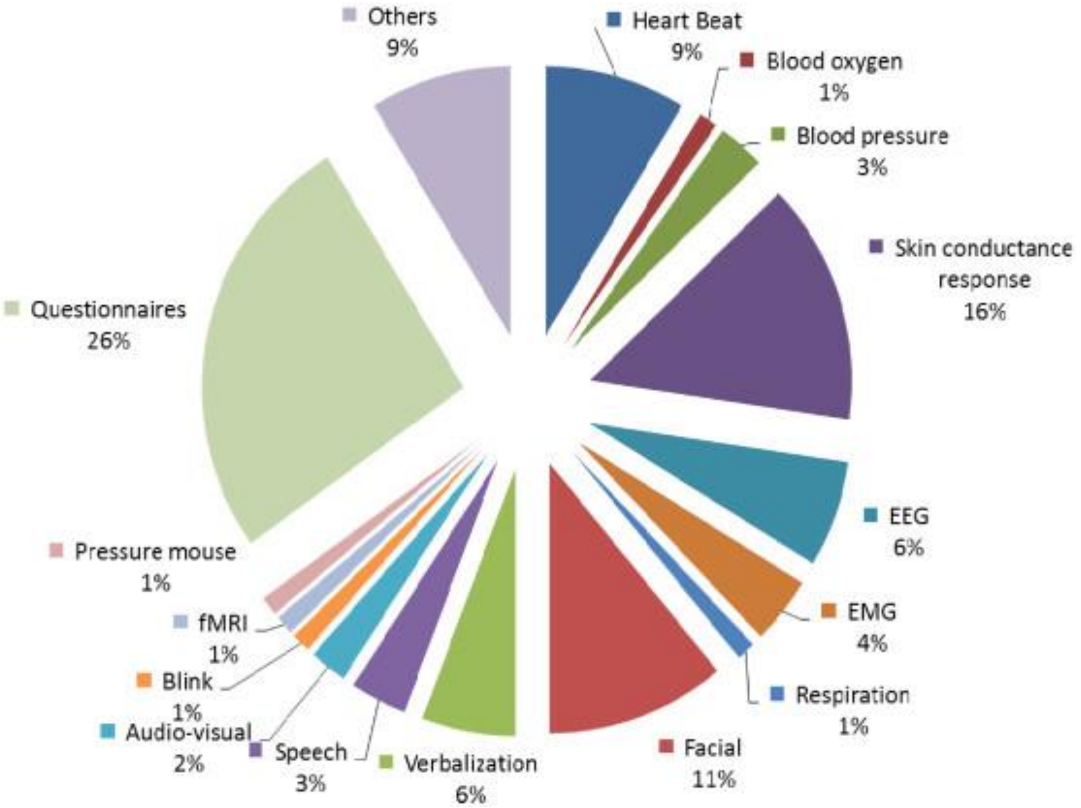


Figure 1 Distribution of data sources used in related works analyzed in [246].

Regarding the data processing of these data sources proposed, we can find two different kind of signals: there are time-frequency based signals, e.g. heart rate (HR) or respiratory rate, based on the repetition of a given pattern in the raw signal recorded (electrocardiogram or respiratory volume variation), and signals that do not present a pattern on the raw signal recorded, e.g. skin temperature or galvanic skin resistance (GSR).When reviewing literature, we can see that most common data sources used in related works are heart rate and galvanic skin conductance. Some works rely on the use of one of those single signals: heart rate [4,14,53,186] or skin conductance [63,96].

As aforementioned, heart rate is a time-frequency based signal, so two different kind of features are commonly extracted from that signal: direct heart rate (i.e. number of heart contractions per unit of time) based features (e.g. mean, standard deviation, etc.) [51,148,184,242] and heart rate variability (i.e. the variation in the interval between heartbeats) based features related to power spectral density (e.g. the different frequencies, high-low frequency ratio, etc.) [14,51,125,148,184,186,242] as this variability often used as a noninvasive test of integrated neurocardiac function, because it can help distinguish sympathetic from parasympathetic regulation of the sinoatrial node [147].

Electrodermal activity is another psychophysiological phenomena widely used as data source as it has been commonly used to distinguish between basic emotions [54] or to measure physiological arousal [99]. As the signal given by this data source does not present a pattern on the data recorded, works usually generate statistic indicators from the data, such as mean, maximum and minimum values, variance, etc. [44,63,96,113,125,184,242,249,251].

A similar processing to the electrodermal activity based data is given to skin temperature data. Some works use the skin temperature as data source in emotion recognition [49,125,249,251], usually generating different statistics from the raw skin temperature data recorded.

When using respiration as a data source, works usually extract the breath rate from the tidal volume (i.e. the normal volume of air displaced between normal inhalation and exhalation when extra effort is not applied) measured. Some works using this data source include [61,125,125,148,184,249]. Other more specific features rarely taken into account could be the end-tidal fractional CO2 concentration [61].

Other psychophysiological and behavioral based data sources that are not used in this research but can be found in affective state detection works include: pupil diameter and other eye tracking based features[107,251], facial expressions[99,204], electromyography [44,249], electroencephalography [95,155,249], electrooculography [249].

It has to be evaluated, due to the nature of this data, the devices required for the collection of the data. Some works rely on expensive devices from well-known brands in laboratory-oriented physiological acquisition devices market to collect that data with a high level of detail [88,125]. Other works, rely on the development of their own data acquisition hardware by means of open hardware solutions like Arduino [49,89]. Another growing tendency is using publicly available wearable devices in order to collect some physiological signals [51,57,97]. Although the use of wearable devices may sound inappropriate in comparison to the traditional laboratory data acquisition devices, there are works that evaluate the cons of using more affordable devices, viewing that the use of wearable devices may also be appropriate [184]. Nevertheless, new ways of collecting some of the physiological signals proposed are being more commonly used, for instance, heart rate traditionally was collected by means of electrodes attached to the body [128], while nowadays, most wearable devices can detect heart rate or even breath rates by means of photoplethysmography (PPG) (a low-cost and non-invasive technique for measuring the cardiovascular blood volume pulse (BVP) through variations in transmitted or reflected light) [129,141,148,233] or new ways to detect some signals as hear rate or breathing without using any contact device are being evaluated [83,148,216].

2.1.2. Computer interaction devices

The literature review shows that over the last few years, affective state detection increasingly began to study the potential of keyboard and mouse interaction devices as information sources. These data sources provide valuable information about the behavioral information of the user while interacting with the computer without needing to incorporate additional hardware to a traditional desktop pc setup. Nevertheless, some of these interactions can be performed by other means depending on the device proposed to be used, as mouse cursor can be controlled by means of a trackpad (device that has less related research works with affective purposes [152]) in case a laptop is being used or new different ways of interaction with devices are arising as touchscreen (interaction method that is increasingly being used in affective computing research [26,84,86]) in case a smartphone, tablet or touchscreen controlled computer is being used. For this reason, we have focused this section on studies that address specific traditional keyboard and/or mouse modeling issues used for affective state detection.

The literature review of this section is reported as follows: section 2.1.2.a describes related works that propose keyboard as the main data source for affective state detection, and section 2.1.2.b includes works which use the mouse as the main data source. After that, works combining both mouse and keyboard features are discussed. Once the different works have been identified, a subsection is included to compile and structure the different features proposed in the literature.

2.1.2.a. ***Keystroke analysis***

The keyboard is one of the most common, less intrusive and less expensive data sources employed in sensor-free approaches to affective state detection. Keyboard interactions have been traditionally used for biometric purposes (aiming to identify users from their unique way of typing) [5,253], but there is evidence that models based on keystrokes dynamics generated for biometric purposes exhibit instabilities due to transient factors such as emotion, stress, drowsiness, etc. [123,158,253]. These instabilities in the user model, caused by emotions, have triggered some studies analyzing keystroke analytics for affective detection purposes [28,237].

Keyboard interactions are commonly recorded as a series of key press and key release events (that is, a sequence in which the user presses a key on the keyboard and then releases it). To create a typing model of the user, different features from those press and release events are generated. In the literature we can find keystroke latency based features, measuring the time interval between two keystrokes [27,42,116,117,131,136,137,199,227,228,238]. Keystroke duration is also a common feature seen in research; this feature is measured as the time (from the key press to the key release events) it takes to perform each keystroke [42,122,131,199,227,238]. Also, some studies take into account the typing speed, as the number of keystrokes per time unit [79,116,117,136,137,228]. Instead of generating the features from all the keystrokes, some studies aggregate those keystrokes in groups of 2 and 3 keystrokes (called digraphs and trigraphs respectively) and generate the features from these combinations including overlapping key press events [76,122,123,199].

Other studies focus on the use of certain keys when typing to generate predictive features. Some of these keys include error related keys (i.e. backspace and delete) [27,116,117,122,123,136,137,199,227,228], style related keys (e.g. capitalization keys) [42,228] or other keys such as space bar, enter key, arrow keys, etc. [122,123,227]. Other features identified in the literature include idle time and pause lengths between the key events [27,79], verbosity or number of different keys used during an interval [27] or features related to the position of the keys on the keyboard, such as hand or finger used for typing according to the position of the keys on the keyboard, stylometry (i.e., the measurement of linguistic “style”, used in authorship attribution and in establishing genre shifts within the work of a single author). Finally some studies ask their participants if they are visual or touch typists, that is if they look at their hands or the screen when typing [42].

It should be noted that the use of keyboard as data source may impact on different methodological issues, such as the elicitation method or task proposed in the data collecting (as it is needed the participant to interact with a keyboard, so the task should require interacting with a keyboard). Some data sources may also be impacted with the use of keyboard, such as GSR sensor, which is placed in the fingers of the participants, so it could impact on the typing performance of the participant (as well as the typing movements may induce noise in the signals collected by the GSR sensor or misplace the sensors).

2.1.2.b. ***Mouse movement analysis***

Mouse interaction has not been so widely applied in affective state detection. However, the methodological approach employed with this data source is similar to that was found in our keystroke analysis: i) recording interaction events (traditionally mouse cursor movements and mouse clicks), ii) grouping them and iii) generating features to create a model that varies according to the affective state of the user. Similar modeling issues associated with keyboard data sources appear when modeling mouse interactions such as user skill level (which can be influenced by the device, as the mouse can be a physical device or a track pad) or stress [101].

Regarding the different mouse interaction modeling approaches found in the literature, several open points have been identified. For instance, when processing mouse movements in order to extract features, many works split the raw data recorded (typically coordinates and timestamps) into small time windows, attempting to adapt these time windows to what could be considered independent mouse movements. While keystroke analysis focuses on modeling of keystrokes (i.e., the press and release events of the same key), when it comes to mouse movements there is no clear definition of what a mouse movement is. When trying to identify what a mouse movement is, different studies propose different points of view: some researchers split the data (according to what they consider a mouse movement) every time the cursor has covered a given distance. For example, [171] splits the data every time the cursor covers 30 pixels. In contrast, other researchers split the data when a pause is found, i.e. there is a period over a given threshold in which no mouse event is registered. The work of [126] is an example of this approach, as [126] splits the data when a pause over 0.5 seconds is found. Other works rely on the speed of the mouse, considering, for example, the end of a movement when the mouse movement speed drops to zero [81]. Finally, there are works that generate the mouse interaction features regardless the mouse movements performed by the participant, such as [227], where all the interactions are split and processed into 5 second time windows.

Another open point found is the features to generate from the mouse interactions. When looking at the different features proposed in literature, some common approaches in generating features can be clearly found. Mouse movement related features include straight distance-based features [81][81], precision features (which refer to the relation between the distance in the location of two events and the actual mouse path covered between those two events) [171], covered distance features [81,100,101,126,227] and relocation-related features of mouse along the screen (i.e., along x and y axis) [126]. Another kind of movement-related feature commonly found in some studies is speed (mouse path length divided by time) [81,101,126,136,137,171,236], with some variations in adjusted speed (actual mouse path length between two button clicks divided by shortest path, and then divided by proposed time window) [171] or instantaneous speed in different trajectory points [126]. Other movement based features include acceleration and instantaneous acceleration [81,126]. Regarding the trajectory described by the cursor, absolute direction is used in [171] (proposing directions such as north, northeast, east, etc.) and angle-related features in [81,126,171]. Some studies examine other mouse interactions such as click frequency [81,116,117,136,137,227] and scroll frequency [227]. Other related work generates features from the periods of mouse inactivity (e.g., total time of inactivity, number of pauses, etc.) [126,136,137,236].

As happens with the keystroke analysis, the use of mouse as data source may arise some dependencies. First, in case some other data source is attached to the hand as it may impact on the regular use of the mouse (as could happen with the use of the GSR sensor attached to some fingers). Another point to have into account when proposing mouse as data source is how the mouse interaction data can be determined by the task proposed. Many interactions to be carried out are delimited by the task and graphical user interface design (e.g. distances, clicks, or even speed if time limits are included in the task).As a result of this review, it can be also seen that most research using mouse movement analysis in affective states detection rely on different sets of features from interaction data. There is a lack of standardization in some basic concepts, such as "mouse movement", which is defined in different ways (straight lines of 30 pixel length in [171], or a concatenation of coordinate changes in the cursor with no more than 0.5 seconds between them [126]). Some studies pointed out the impact of using context information in predictive models [171] or considering prevalence of an emotion and the persistence of that emotion over time after the presentation of a given stimuli [126]. Due to the differences found in the review of research related to mouse analysis, it can also be concluded here that there are no consistent general approaches nor a clear reference framework of features. Lastly, the implications of this present study may likely benefit real-world students.

2.1.2.c. *Computer interaction devices feature overview*

Table 1 includes a survey of the different features extracted from the proposed interaction devices in different related works:

Data source	Feature type	Application of feature type	Refences
Keyboard	Keystroke latency	Keystroke latency: time interval between the key release of the first keystroke and the key press of the following keystroke	[27,42,116,117,131,136,137,199,227,228,238]
		N-graph grouped keystroke latency: keystroke latency calculated separately between first and second (digraphs and trigraphs), and second and third (trigraphs) keystrokes of a n-graph.	[76,122,123,199]
	Keystroke duration	Keystroke duration or dwell time: time interval between a key press event and the release event of that key	[42,122,131,199,227,238]
		N-graph grouped keystroke duration: keystroke duration calculated separately for first and second (digraphs and trigraphs), and third (trigraphs) keystrokes of the n-graph.	[76,122,199]
		N-graph total duration: duration of a n-graph (digraph or trigraph) from 1st key down to last key up	[76,122,199]
	Typing speed	Typing speed: total number of keystrokes or words per unit of time (minute, second, etc.)	[116,117,136,137,228]
	Style related features	Capitalization Rate: Capital to lowercase character ratio	[42,228]
	Key related features	Frequency of enter or spacebar keystrokes	[122,123,227]
		Frequency of error related keys: use of backspace key and delete key	[116,117,122,123,136,137,199,227,228]
		Verbosity: number of different keys used	[27]
		Common/rare consonant and vowel frequency related features	[42]
	Other features	Sequence: a list of consecutive keystrokes	[228]
		Hand-based: hand used related features according to the hand that would be used to type each key based on “touch-typing” norms	[42]
		Finger-based: features related to the finger that would be used to type each key based on “touch-typing” norms	[42]
		Keyboard row: features regarding key location on keyboard	[42]
		Idle time, pause related features: number of time periods over a given threshold with no interactions	[27,79]
		N-graph number of events: number of key events that were part of the n-graph (digraph or trigraph).	[76,122,199]
Mouse	Distance features	Precision: relation between distance of two events location (mouse movement start and end, button clicks, etc.) and actual mouse path covered between those two events.	[171,199]

Data source	Feature type	Application of feature type	Refences
		Distance: total distance covered by mouse cursor	[81,100,101,126,227]
		Relocation related features of mouse along the screen (i.e., along x and y axis)	[126]
	Speed	Speed: mouse path length divided by time (overall speed, between two button clicks, between pauses, etc.).	[101,126,137,171]
		Adjusted speed: actual mouse path length between two button clicks divided by shortest path, and then divided by task completion time.	[171]
		Instantaneous speed of mouse: speed of mouse in different points of the trajectory.	[126]
	Acceleration	Acceleration: speed change over time	[126]
		Instantaneous acceleration: speed changes of the mouse in different points of the trajectory.	[126]
	Direction/angle	Direction: number of mouse movements in a particular direction (In [171] the directions proposed were north, northeast, east, etc.).	[171]
		Angle features: angles described by the mouse trajectory (In [171], where angles are grouped from 0 to 180 degrees by 10-degree step, while in [126] average and sum of angles are calculated).	[126,171]
	Mouse elements interaction	Left/right click frequency	[137,199]
		Scroll use	[199]
	Pause	Pause features: generated from mouse inactivity times (total time, number of occurrences, average time, etc.).	[126,137]
		sensitive pause features: generated from mouse inactivity times over a given threshold (0.5 seconds in [126]).	[126]

Table 1 Computer interaction devices (i.e. keyboard and mouse) features used in related works.

2.1.3. Multimodal approaches

Many of the works previously introduced rely on the use of a single data source, while some of them combine several of them. While some works have compared the results of performing affective state detection by means of single data sources with combination of data sources [148] (with favorable results for the combination of all the data sources), that is something that is not commonly evaluated as related works usually combine all the data sources proposed without evaluating the results of single data source approaches.

Although some works using physiological signals rely on a single data source, being most common unimodal physiological approaches based on hear rate related features

[14,53,186] or electrodermal activity related features [63,96], most works combine different physiological signals [49,88,99,125,242,249].

When looking interaction devices most works rely on the use of a single data source (keyboard [42,79,123,238] or mouse [81,100,126,236]). When looking at works that combine both keyboard and mouse, there are not so many woks that evaluated that combination [117,136,137,227]. Regarding the combination of these interaction data sources with physiological signals, we can find some approaches using physiological devices [116], modified interaction devices (as a mouse capable of measuring galvanic skin resistance [82] or wearable devices [202].

Past work has combined both data sources in affective state detection. These have followed approaches that may take into account not only issues related to each data source separately, but also features that may appear from the combination of both, such as the ratio of interaction with each data source, pauses when switching from one source to another, etc. Nevertheless, little research has been done in comparing unimodal and multimodal approaches. That is one of the most clear open points identified in the literature regarding the data sources to be used in affective state detection, so that will be one of the spots where our research focus has to be set.

2.2. Data analysis technique

As most works aim to detect affective states from the data collected from the different data sources proposed, they use machine learning techniques in order to perform that detection task. Because of this, that is the choice taken in this work (as was already mentioned in section 1.3). There are some works not using machine learning techniques, which main goal is evaluating the significance of the impact of some variables from the data sources in the affective state, using analyses such as correlation [123], t-test [237], ANOVA [61,95,126,131,136] and MANOVA [43,116,137].

Regarding the works aiming detect affective states, machine learning is the best way to carry out that detection due to the amounts of data collected from different data sources. We can find many different choices between the data mining algorithms and techniques used in related works: K-Means [227], K-Nearest Neighbors (K-NN) [44,96,122,125,227,238,249], Bagging [227], Naive Bayes [27,42,113,122,125,148,249,251], IB1 [227], KStar [227], Random Committee [227], Random Forest [27,89,125,171,227,239,239,248], Bayesian networks [27,113,122], Random Tree classifier [27,79,159,227], Support-Vector Machines (SVM) [27,42,44,49,96,113,125,148,155,159,171,184,248,251], Generalized Linear Mixed Models (GLMM) [4], Deep learning techniques [249], logistic regression [117,159,171,249], Least Square SVM (LSSVM) [249], Neural network [44,122,125,159], RIPPER [125], C4.5 (or J48) tree [27,79,125,159,171,236,239,240], OneR [27,76,125,186], Decision Table [27], REPTree [27] or decision tree classifiers [42,113,122,122,239,251]. As we can see from the techniques mentioned, K-NN, Naïve Bayes, Support Vector Machines and C4.5 techniques are the most used in the selected related works.

Other thing to take into account is the use of preprocessing techniques used in some of these studies. We can find works using Principal Component Analysis [44,238,249], Forward Feature Selection (FFS) [125], Minimum Redundancy – Maximum Relevance (mRMR) [125], ReliefF [125], Information gain (IG) [125], Chi-squared (Chi2) [125], Student’s t-test [238], ANOVA [44] or down-sampling [27,76,238]. In this sense, we can see how related works traditionally do not provide detailed information on how the data is processed further than the algorithms used for the model generation. This preprocessing point and its lack of detail in most of related works is one of the issues we aim to address in this work (in stage 2), introducing some of the techniques mentioned in this paragraph as a methodological variable, evaluating the impact of using them.

Figure 2 includes a simple data processing flow (which usually consists on preprocessing the data collected before the model generation) summarizing the most common predictive algorithms and data preprocessing techniques found in the related works.

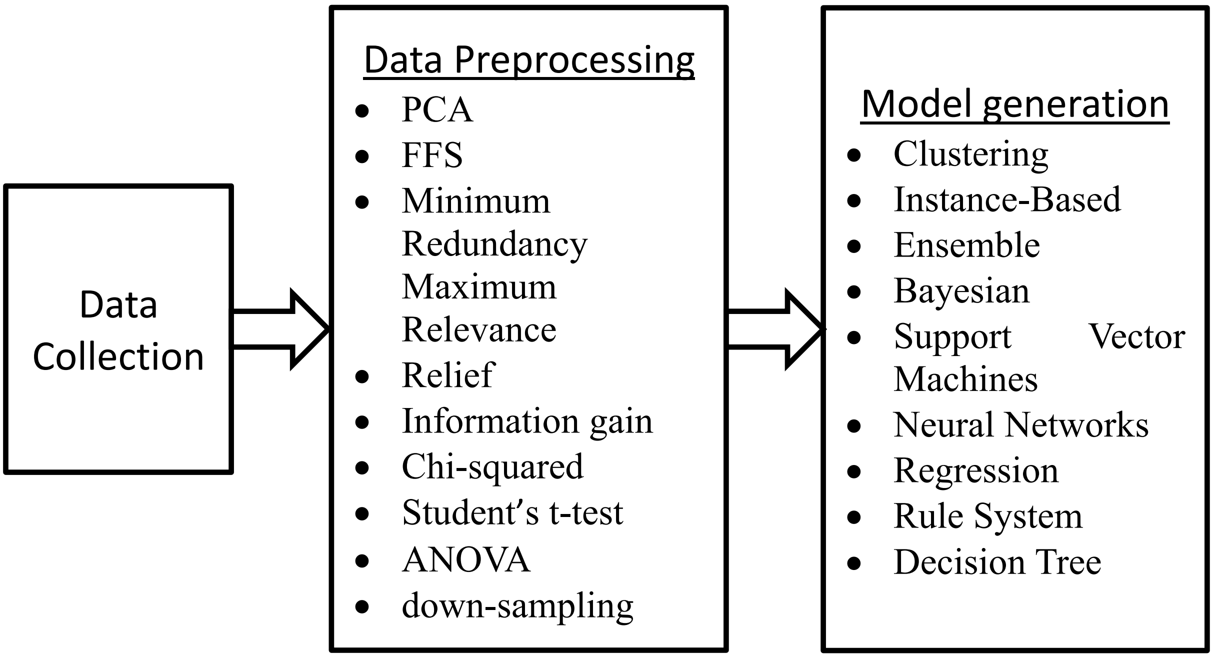


Figure 2. Data preprocessing and processing techniques identified in related works

Some works also point the volume of data used in their research. Given the lack of individual user data in educational contexts [188], one critical common problem is high dimensionality, that is, having many more features to describe users´ interactions than the number of available instances [31,122]. Many of these studies deal with relatively small datasets. This is an issue that can have an strong impact con the experimental methodology, as in [28] authors addressed the shortage of data representing certain affective states, modeling only the states that comprised the majority of the affect labels (thus avoiding building models based on few observations). It is not hard to find works with an amount of participants ranging from 3 to 9 [63,79,122,123,186] to works reporting more than 300 participants [42,248]. Regarding the number of data instances used, some works use less than 100 data instances for the data analysis [63,96,100,148].

Table 2 includes a survey on the data sources, data analysis techniques, top results (when performing affective state detection) and number of participants and data instances used in the works described in this section.

Ref.	Year	Data sources	Algorithms	Best accuracy	No. of participants	Data instances
[81]	2017	Mouse interactions, context info	Conditional probability	First experiment (Emotional labeling): 68.43% accuracy Second experiment (web search): 70.15% accuracy	15 participants aged 21 to 31; 7 females, 8 males;	First experiment: 3645 annotations (15 users * 243 annotations/user) Second experiment: 276 web search questions (46 total sessions * 6 questions per session)
[136]	2017	Mouse interactions, keyboard interactions, typo mistakes	ANOVA, Spearman Correlation test, MANOVA		190 participants in total (only 162 completed the task)	972 total instances (6 data instances per participant)
[227]	2017	Mouse and keyboard	K-means, KNN, Bagging, IB1, KStar, RandomCommittee, RandomForest, RandomTree classifier	69% accuracy	35 participants 28 male and 7 female	
[123]	2018	Keyboard	Correlations		9 participants	207 samples (from 8 to 47 samples per participant)
[184]	2017	Physiological signals: cardiac and electrodermal (EDA) activities Devices used: Biopac MP150 and Empatica E4 wristband	Support Vector Machine (SVM), and supervised learning algorithms	66% accuracy for valence level 70% accuracy for arousal level	19 participants 12 female and 7 male; average age 33.89 years ± 8.62	971 samples (20 not labeled)
[14]	2017	Electrocardiography sensor Device used: SHIELD-EKG-EMG for Arduino			10 participants 1 female, 9 male; age between 32 and 47 years old	2 courses per participant
[53]	2017	Heart rate variability sensor Device used: PolyG-A by Laxtha			19–35 years old participants	

Ref.	Year	Data sources	Algorithms	Best accuracy	No. of participants	Data instances
[4]	2016	Heart rate Device used: Fingertip Pulse Oximeter MD300C21 by Beijing Choice Electronic Tech	Generalized linear mixed models (GLMMs)		24 participants 11 males and 13 females; ages ranged between 20 and 38 years.	144 total instances (calculated from 6 instances per participant)
[249]	2017	Electroencephalography, electrooculography, electromyography, skin temperature, galvanic skin response, blood volume pressure and respiration	Multiple-fusion-layer stacked autoencoders (SAEs), KNN, Logistic Regression, least square support vector machine, Naïve bayes, Principal Component Analysis + Naïve bayes, Laplacian eigen-maps + Naïve bayes, and neighbor preserving embedding + Naïve bayes	Arousal (accuracy)=0 .8418 Arousal (F1)=0 .7798 Valence (accuracy)=0 .8304 Valence (F1)=0 .7950	32 participants 16 male, 16 female; 19–37 years old; mean age = 26.9	40 videos for each participant
[88]	2016	Electrocardiography and galvanic skin response Device used: a 16-channel PowerLab by AD Instruments	wavelets, probabilistic neural network	97.90% and 97.20% accuracies were reached for GSR and ECG signals (sigma=0.01)	11 participants 11 female university students (age range: 22.73±1.68 years old)	12 instances per participant
[49]	2017	Heart rate, galvanic skin response, skin temperature Devices used: Arduino Uno, Raspberry Pi B+, Heart Beat Sensor, GSR Sensor, Temperature Sensor	Support Vector Machines			

Ref.	Year	Data sources	Algorithms	Best accuracy	No. of participants	Data instances
[125]	2014	skin conductance, electrocardiography, respiration and skin temperature Device used: BIOPAC MP 150	kNN, SVM, RF, MLP (multilayerPerc), RIPPER, C4.5, NB	60,3% accuracy	14 female participants age: avg 20.3yrs, std 0.8	686 total instances labeled as: 226 "other", 78 "happiness", 38 "fear", 99 "disgust", 89 "sadness", 156 "neutral"
[99]	2015	Electrodermal activity and facial expression		Agreement	67 participants 82.1% female, mean age of 21.00 (SD = 1.90), mean GPA of 3.14 (SD = 0.69), 74.60% Caucasian	5 instances per participant
[148]	2014	Heart rate, Breath and Heart rate variability Devices used: a digital single-lens reflex (DSLR) camera with a standard Zuiko 50mm lens, Flexcomp Infiniti by Thought Technology Ltd.	NB and SVM	85% accuracy	10 participants 7 females, 3 males; ages between 18-30	2 instances per participant
[242]	2017	Electrocardiography and skin conductance Device used: Procomp Infiniti by Thought Technology Ltd.	Multi-Label learning	66% accuracy for valence (F1=0.63) and 81% accuracy for arousal (F1=0.75)	25 male participants age between 20 and 33 years old	100 data instances in total
[51]	2017	Heart rate and learner's performance Device used: apple watch			20 participants (10 control group and 10 experimental group) 11 male and 9 female; ages ranging from 18 to 21 years old.	The system provided 10 English grammar exercises with 40 multiple-choice questions per exercise. Students were asked to complete as many exercises as possible.

Ref.	Year	Data sources	Algorithms	Best accuracy	No. of participants	Data instances
[89]	2017	Electrocardiography, Electromyography (on 4 facial muscles), Galvanic Skin response and respiration rate. Device used: Arduino Due, shield-EKG-EMG by Olimex	Random Forest		10 males between age range between 18 and 38 years old	47464 instances in total
[238]	2017	Keyboard	KNN	91.21% accuracy (0.8241 kappa; 0.9089 AUC; 0.9215 F)	53	985 instances in total (Avg 17 per participant)
[27]	2013	keyboard	J48, NaïveBayes, BayesNet, SMO, DecisionTable, OneR, RandomForest, RandomTree, and REPTree.	87.0% accuracy (kappa=0.374) for boredom and engagement classification; 56.3% accuracy (kappa=0.171) for three-way boredom-engagement-neutral discrimination	44 participants 68% female; mean age of 19.9 years; 45% Caucasians, 52% African Americans, and 3% "Other"	
[79]	2012	keyboard	C4.5, Random tree and BF tree	95.79% accuracy (Kappa=0.9262)	6 participants 4 male and 2 female	1502 instances in total
[117]	2008	keyboard, mouse	Logistic regression analysis		26 participants 24 male, 2 female; mean age was 27 years with SD of 3; age range from 22 to 34 years.	
[116]	2013	keyboard, mouse, GSR	MANOVA		16 male participants mean age of 26 years (SD = 3.1)	60 instances per participant

Ref.	Year	Data sources	Algorithms	Best accuracy	No. of participants	Data instances
[42]	2015	keyboard	NaiveBayes, AdaBoost with single split decision trees, SVM with an RBF Kernel, and SVM with a Linear Kernel.	72% accuracy	486 41.3% female,56.4% male; 79.7% native English speakers, 17.0% non-native English speakers; 88.3% right-handed, 9.1%left-handed	10 - 12 instances per participant
[28]	2013	keyboard		87.0% accuracy (kappa=0.374)	44 participants	3 instances per participant
[78]	2011	keyboard	C4,5	87.8% accuracy (kappa= 0,76) with the sadness model	12 participants 10 male, 2 female; age between 24-34 (mean=28.5, s.d.=2.7)	From 51 to 219 instances per participant (mean=94.1, s.d.= 52.7)
[122]	2015	Keyboard	decision trees, neural networks, k nearest neighbors, naive Bayes, AdaBoost, rotation forest, Bayesian networks.	81,25% accuracy	9 participants 2 female, 7 male	
[131]	2014	Keyboard	ANOVA		27 participants age between 19 and 27 (M = 21.5, SD = 2.3)	1620 total instances (60 instances per participant)
[137]	2014	keyboard, mouse and performance	MANOVA		60 participants (after 17 participant data were considered invalid samples) 90% male; age between 18 - 24 years old	
[228]	2013	keyboard				
[237]	2013	keyboard	t-test		15 participants 10 male, 5 female; mean age of 23.4 years old, std = 1.45	80 instances per participant

Ref.	Year	Data sources	Algorithms	Best accuracy	No. of participants	Data instances
[100]	2016	mouse	descriptive statistics		i) 65 participants: 33 in control group (48% male) and 32 in negative condition group (46% male); ii) 126 participants: 62 in control group (62% male) and 64 in negative condition group (64% male); iii) 80 participants (40 germans and 40 chinese)	i) 1 instance per participant, ii) 1 instance per participant and iii) 5 instances per participant
[171]	2015	mouse	Logistic regression, Support Vector Machine, Random Forest, and C4.5.	94.61% accuracy (0,946 F-Score, 0,946 ROC)	44 participants	1056 total instances (44 participants x 24 instances per participant)
[248]	2013	mouse	support vector regression (polynomial kernel with degree 3)	0,63 Correlation coefficients between predicted and observed state anxiety scores	367 participants 234 participants (137 female, 97 male) participated in the initial feature selection experiment. Additional 133 participants (75 female, 58 male) took part in the evaluation experiment.	96 instances per participant
[126]	2014	mouse	ANOVA		14 participants	
[236]	2011	mouse	C4.5 (DTC)	97.24% accuracy	136 participants	
[230]	2012	mouse	regression model		131 participants	124 total instances
[240]	2013	Heart rate and galvanic skin response Device used: The BioHarness pulsimeter by Zephyr	C4,5	95-99% accuracy	12 participants 7 male and 5 female	

Ref.	Year	Data sources	Algorithms	Best accuracy	No. of participants	Data instances
[43]	2013	speech, electrocardiography and personality traits	MANOVA		40 participants	30 instances per participant
[186]	2010	Electrocardiography	1-R rule (Decision Tree Classifier)	81,2% accuracy	6 participants	
[61]	2001	Video analysis, Oxymetry, FetCO2 level, respiratory inductive plethysmograph	ANOVA		40 participants	8 instances per participant
[251]	2006	Blood volume pressure, galvanic skin response, skin temperature and pupil diameter Devices used: GSR2 module by Thought Technology LTD (to collect the Galvanic Skin Response), photoplethysmography (to measure the blood volume), an LM34 IC (to measure the subject's skin temperature), NI DAQPad-6020E for USB by National Instrumentation Corp (to digitized signals) and the ASL-504 eye gaze tracking system to get an accurate and continuous pupil diameter signal	NB, DTC and SVM	90,1% accuracy	32 participants	225 instances per participant
[96]	2010	Galvanic skin response	SVM and KNN	97,06% accuracy	17 participants	2 data instances per participant

Ref.	Year	Data sources	Algorithms	Best accuracy	No. of participants	Data instances
[44]	2010	electromyography and galvanic skin response	KNN, SVM, ANN	61,3% accuracy	24 participants 20 females; average age 43 years; recordings of three subjects were discarded	8 data instances per participant
[63]	2014	GSR	One-tailed test		3 participants	15 data instances per participant
[113]	2014	electroencephalography, galvanic skin response and blood volume pressure	BN, NB, DTC and SVM		44 participants 31 males; age between 19 and 52 (M = 28.61 ± 8.40)	1848 total instances (42 data instances per participant)
[155]	2014	Galvanic skin response, electroencephalography	SVM	51% accuracy	12 participants	120 data instances per participant
[95]	2012	electrocardiography and electroencephalography	ANOVA		16 participants 6 females, 10 males; mean age of 26.4 years (ranging from 21 to 35 years)	60 data instances per participant
[239]	2016	keyboard and mouse	J48, Decision Tree, and Random Forest	74,28% accuracy (Kappa=0568)	55 participants	
[159]	2014	keyboard and text pattern	simple logistics, SMO, MultilayerPerceptron, Random Tree, J48 and BF Tree	87% accuracy	25 participants (15 to 40 year old, about 45% female) in fixed text experiment	

Table 2 Data sources, data analysis technique and top results reported (with prediction algorithms) in related works

As we can see in Table 2, best accuracy values vary from 51% to 99%. Regarding the number of participants, we can find works with more than 400 participants and works with 3 participants. When evaluating the data processing performed, we can see also a wide variety of techniques. Some of the most common As aforementioned, due to the huge variety of approaches proposed, each one different from the other, it is impossible to evaluate which methodological issues lead to better results due to the impossibility to compare these related works.

2.3. Labeling

As the proposed approach in this research relies on the use of supervised learning techniques (like most of the studied works), the data obtained from our experiments should be carefully labeled in order to have an accurate dataset to train the different algorithms to be used. The importance of the labeling on the predicting task is crucial in order to get a lifelike model and avoid noisy data. In particular, the labeling characteristics are one of the most controversial and open points nowadays in affective states detection due to the nature of the problem addressed. There are many factors to take into account when choosing a labeling approach, having each of the possibilities pros and cons, as discussed below. In particular, they have an impact on the data mining algorithms that can be used. A description of the different aspects that have been identified on the emotional labeling approaches reviewed is presented in this section. For this analysis, the different labeling aspects considered have been: i) labeling format, ii) subject to label the data, iii) frequency of the labeling, iv) time of the labelling, and v) labelling presentation. Each of them is discussed next.

2.3.1. Labeling format

The format with which the emotional labels will be obtained is a point that has to be wisely chosen, as this may suppose some barriers in future processing (such as limit the algorithms to be used as some algorithms can handle only numerical values or only categorical values). Some approaches can be used to transform some labeling formats into others, but only if the original labeling format allows that conversion. But not only the data analysis algorithm to be used are to be taken into account choosing the labeling format, as many other aspects may have an impact of the appropriateness of the format chosen. For instance, some works suggest that, depending on the individual to use the emotional model chosen, one choice or another can be better (e.g. [23] suggests that a discrete emotional model would be more appropriate for individuals who focus both on pleasantness and on their level of subjective arousal when labelling their subjective emotional experiences, while a dimensional model would be better to be used with individuals who focus mainly on the pleasantness or unpleasantness of their subjective emotional experiences). Nevertheless, after more than a century since the first psychological models proposed, there is still today a lack of agreement on the choice of a given emotional model in scientist who study emotion [72].

2.3.1.a. ***Categorical labeling***

When dealing with emotions in our lives, we usually adopt a categorical approach as we usually (try to) name our emotions. Asking people to label their feelings offering them a closed list of emotions [27,49,63,88,130] means dealing with some issues, as commented next.

First, deciding if the users will be able to choose just one or several emotions from the list as they may feel a “pure” emotion or a mix of them (in [240] the focus is on one emotion while other works offer up to 15 emotions [27,76]). In case they are able to choose several emotions from the list, then it has to be decided if they will be able to select the intensity of each one (in [78] the emotional state questionnaire contained 15 questions evaluated in a 5-point Likert scale regarding a user's current emotional state) of the chosen emotions as they may feel a mix of different emotions but in different degrees.

Another relevant issue is which emotions will be on the list as a list of all the known emotions can be huge and make the labeling process slow, and even affect the labeling if looking for the experienced emotion in a huge list of emotions for a long time makes the subject feel frustrated. The EmotionML standard regards the existence of many different approaches [221] and some of them can be found in [257]. In case a list of a limited set of emotions is to be offered, it has to be decided which criteria will be used to choose those emotions, as these criteria may vary from the cultural aspects of the subjects [153] (as in different cultures some emotions can be perceived in a different way or even expressed differently[55]), to the nature of the task labeled (if the aim of the experiment is to elicit "disgusting" emotions, the focus on the list should be on those emotions, but some participant may feel other kinds of emotions that might not be included on the proposed list).

In this regard, we can see how works dealing with affective state detection framed in learning environments usually set the focus on states that are more common in learning [124] such as stress [136], boredom or confusion [79], attention [14], anxiety [51], etc.

Assuming all the subjects know all the emotions appearing on the list and that all the subjects have the same (or a similar) conception of each one of the emotions listed (this is something related to the cultural background of the sample used and the fine grained the emotion list has been made, offering similar emotions slightly different), then it has to be taken into account the way the list is going to be presented (as from the emotions chosen to be shown to the order they are shown may induce subject's response).

It should be noted that following the categorical approach the prediction techniques that can be used to process the data are limited to those that can handle a categorical labeling [52].

Looking at some related works in affective state detection, we can see that some works have used this approach for labeling data [27,28,49,76,88,99,122,227,228].

2.3.1.b. *Dimensional labeling*

In contrast to the categorical models, there are other affective models in order to "evaluate" emotions based on different aspects of the emotion. To do this, first is needed to define an emotion as the combination of different characteristics that define each emotion, so this kind of approaches usually are based on the use of different dimensions to model an emotion.

This dimensional approach appeared in 1897, when Wilhelm Wundt detected "Three such chief directions may be distinguished; we will call them the direction of pleasurable and unpleasable feelings, that of arousing and subduing (exciting and depressing) feelings, and finally that of feelings of strain and relaxation."[247]. And this approach has been followed by others all along last century [166,218,219]. The only dimension that appears commonly in all dimensional proposals is the pleasure vs. displeasure, followed by the intensity of the emotions. Other less common dimensions are: control (or dominance), attention vs. rejection and relaxation vs. tension [234]. These two dimensions are the most common ones in the works reviewed using a dimensional labelling approach in affective computing [43,44,95,117] with some of them using another dimension as dominance [230].

Some of the most common approaches nowadays rely on these dimensions, such as the Russell's "Circumplex Model of Affect" [189], that hypothesizes that the emotional space "could be defined in terms of two orthogonal dimensions, pleasure-displeasure and degree of arousal " (see Figure 3) and the Pleasure-Arousal-Dominance (PAD) emotional state model, that describes the emotional states from three dimensions (Pleasure-Displeasure Scale, Arousal-Nonarousal Scale and Dominance-Submissiveness Scale) [151].

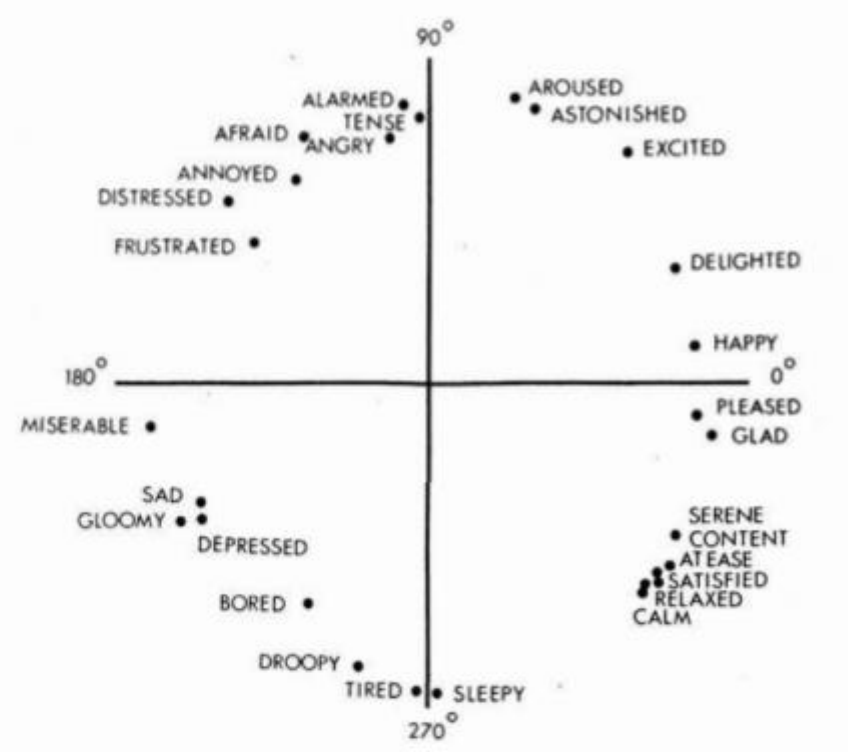


Figure 3 Direct circular scaling coordinates for 28 affect words proposed by Russell [189]

In any case, the dimensions to be used to evaluate the affective state of the subjects have to be well and clearly explained in order to guarantee the labelling really represents the emotions evaluated.

When following this approach, data mining algorithms capable to handle numeric values as class attribute can be used. In addition, as the values can be binned into categories (for instance, if we are interested in predicting certain values range instead the concrete value of each dimension), can also been used categorical prediction algorithms.

Dimensional labelling has been widely used in related affective state detection works, being the most followed approach the circumplex model of affect [43,53,89,100,117,131,184,230,249].

2.3.1.c. *Mixed approaches*

There are also approaches that mixed different kinds of labels, such as the Plutchik wheel of emotions [176] or the Lövheim cube of emotion [140].

The Plutchik wheel of emotions is a model that relies on a categorical list of eight basic bipolar emotions (anger vs. fear, sadness vs. joy, disgust vs. trust, surprise vs. anticipation) that can be displayed as a wheel (2D) or a cone (3D) (see Figure 4), being the cone’s vertical dimension (or the distance to the center of the wheel in the 2D representation) the intensity of the emotion [176]. On the issues to be taken into account when adopting this approach, is to include all the previous points spotted, being aware that depending on the nature of the values to ask for, the algorithms to be used may be limited.

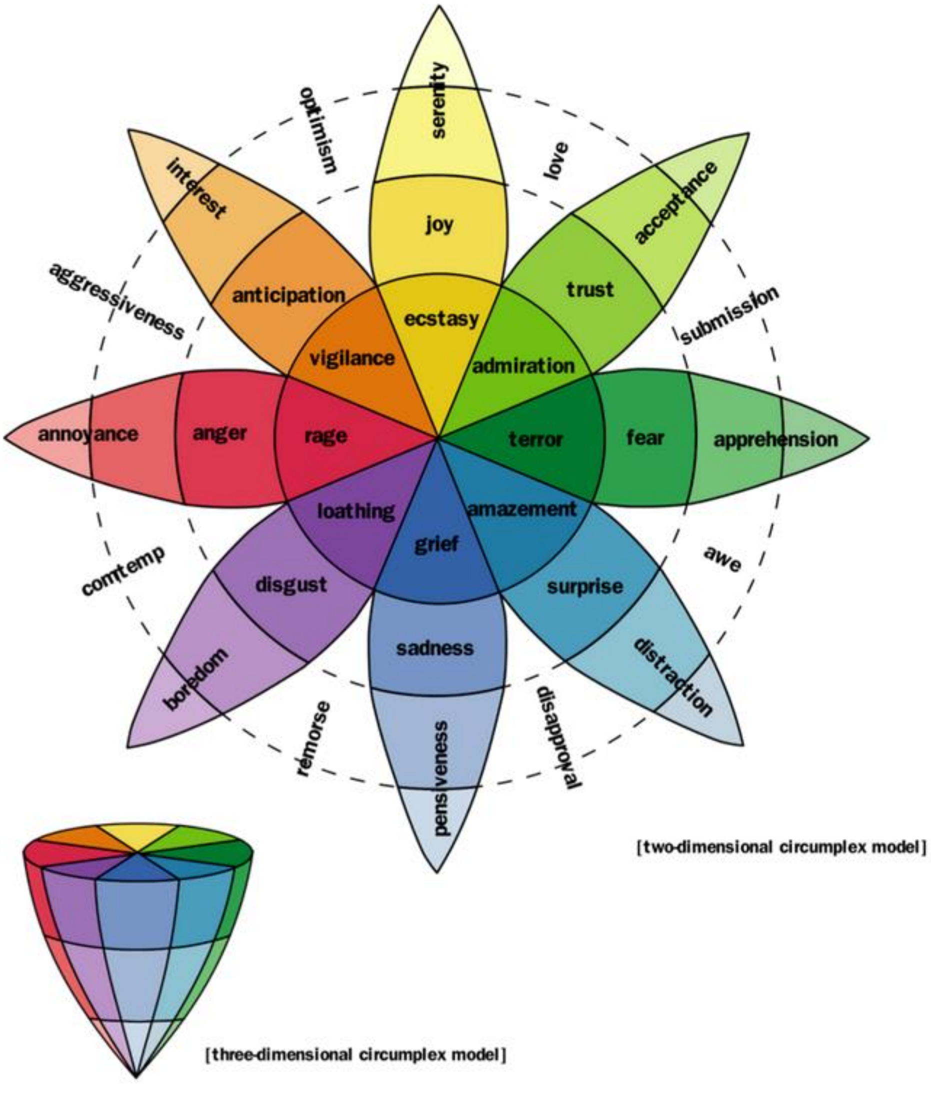


Figure 4 Plutchik's wheel of emotions [176]

The Lövheim cube of emotion [140] is a model that establishes the relations between three monoamine neurotransmitters: serotonin (5-HT, 5-hydroxytryptamine), dopamine (DA) and nor-adrenaline (NE) and eight basic emotions (depicted in Figure 5).

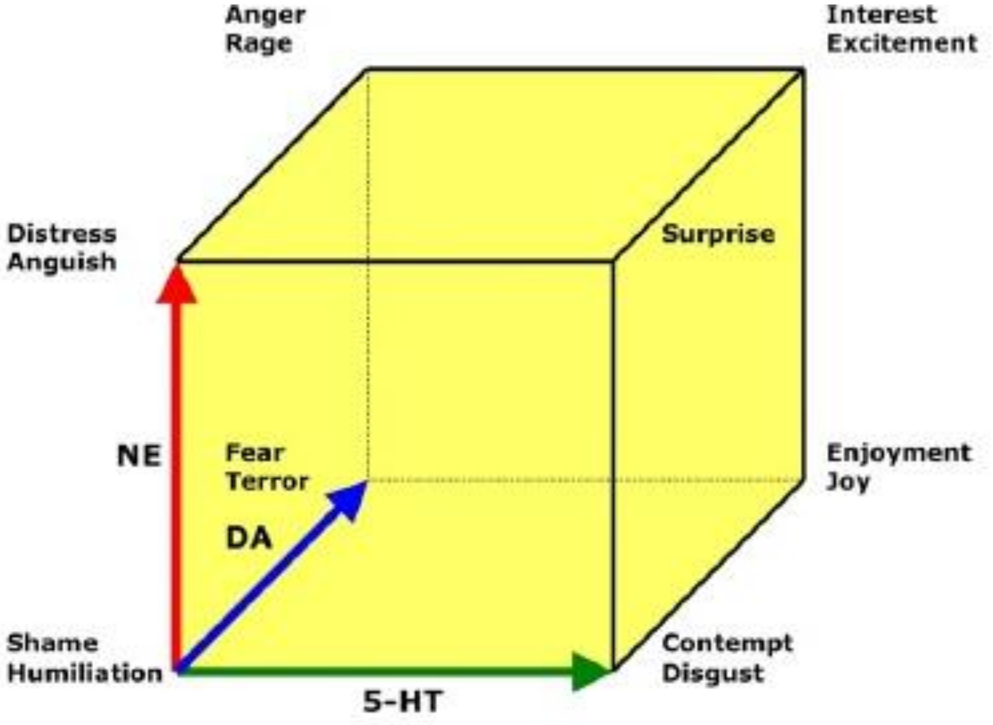


Figure 5 Lövheim cube of emotion [140]

2.3.1.d. *Open labeling*

Another possible approach is to allow the subject to express his or her emotions in a free way, being this: i) writing a list with the emotions felt, ii)describing any dimension of his or her affective state or iii) just explaining how he or she feels without giving any specific emotion name (e.g., “I felt bad”).

This approach brings many problems when processing it to define labels that can be used to extract the values to be predicted. If the subject is asked to provide emotion names, these may be used as the values to be predicted, but before using them, it should be contrasted that the words given by the subject are really emotions (so they should be matched to an emotion list to check that all the words given are emotions in order to avoid the use of personal created terms to name an emotion.

In case the subject is allowed to express himself or herself with no limitations (i.e., free text, regardless if it is captured by typing, talking, etc.) many different approaches can be done after that, being the most plausible a natural language processing one [47] performing sentiment analysis [150]. In this case, we have to set the values to be extracted from the text to be generated (and that will be predicted), returning to the problem addressed previously, as we will generate labels (categorical or numeric) to be predicted.

2.3.1.e. *Conversions*

Some studies have been carried out to see the correlation between categorical labels and dimensional ones [23,70], commonly finding correlations but not with high values. Although converting from a dimensional approach to a categorical one can be feasible, the reverse conversion, from a categorical approach to a dimensional one can be more difficult (more feasible, but still difficult, if the dimensional values are grouped into categories).

From the references studied, in [43] a categorization of images is performed using their affective dimensional labeling to label them into 6 basic emotions. In [242], the numerical dimensional values collected are divided into high and low level. In [227], authors select nine videos, each one corresponding to a given reference emotion, but also assign each one of the videos a category according to the high/low arousal level of the emotion and positive/negative valence of the reference emotion.

2.3.2. Subject to label the data

Other issue to deal with when designing an affective computing experiment is who will be responsible for labeling the data. There are four possible approaches to be followed, which can be compatible with each other.

2.3.2.a. *User*

The most common approach followed in affective computing works is asking the subject to provide the information to label the data. Due to the intrinsic nature of the problem, the most plausible way of knowing what is someone’s feeling is letting them to express it.

A point to face from this approach is that to provide a trustworthy image of one self-feelings, a certain self-awareness capability is needed, which is not common and may influence the emotions labeled [229]. Other aspect to have into account in this case is how other factors as the emotional representation may impact on the comprehension of the phenomena to be labeled and how that representation may be more or less suitable depending on the subject [23].

From the reviewed works, this approach is the most used as can be seen in [53,63,122,130,131,184,227,240,249].

2.3.2.b. *External expert*

Another commonly followed approach is asking an emotional expert (usually a psychologist) to label the emotions experienced by the user during an experiment [79]. Some issues should be clarified before following this approach such as the background the labelers should have (psychologist, experts of the field the experiment is being carried in, people close to the subject, etc.) and the criteria to be followed when labeling the emotions. This last point is crucial if the labeling will be done in different sessions or there will be more than one evaluator. To deal with that, some emotional labeling methodologies have been proposed and evaluated, such as the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) [162], a labeling methodology to be used in real world educational scenarios[32], or the Human Expert Labeling Process (HELP) [17], a methodology to be used with multiple labelers in educational scenarios to label collected data and obtaining agreement among them on same samples of data.

In any case, there is a need to look for the less labeler-dependent variations in labeling. The external expert should have as much information as possible from the experiment in order to be aware of all the possible factors that may trigger an emotion and all the reactions of the subject [79].

2.3.2.c. ***Automatically / context driven***

Another option is supposing that certain contextual events may induce affective states in the user, so a rule system can be created that automatically annotates those events and the emotion to be induced [88,96,100,115,251]. This approach can be adopted when dealing, for instance, with standardized affective stimuli, such as sets of affective images [127] or sounds [40], labeled by a big culturally similar sample.

This approach combined with the personal subjective labeling can also be used to evaluate the level of knowledge the subject has about the evaluating method used (looking for a high conformity between the standardized labeling and the subject one).

2.3.2.d. ***Mixed approach***

Other approaches can arise when combining the above ones. Although these approaches can be carried out in parallel, they can be also performed altogether, for instance, in a post experimental labeling, with the expert helping the subject to label his or her emotions (but trying not to induce any), or contrasting a labeling performed during the experiment by one or both of them (i.e., expert and user).

2.3.3. Frequency of the labeling

When designing an experiment to collect affective data, another point to design carefully is the moment the labels will be taken. As the main aim of an experiment is to collect all the emotions experienced during an experiment the ideal solution would be that one that detects affective states changes in order to ask for a label. This factor will determine important points when processing the data such as the number of registries obtained (the more, the better) from the experiment.

When deciding the frequency of the labeling, it is important to realize that a high frequency may be intrusive, affecting the subject’s emotions (e.g. feeling frustrated when asking too many times for the labeling, making him or her to lose a lot of time in labeling instead of advancing in the current task).

This issue is closely related to the processing of the data when creating the dataset, because usually, the information obtained from the data sources is grouped for each time window and assigned the label obtained at the end of that time window (assuming the emotion reflected on the label has appeared constantly during the time window). Different approaches for frequency labeling are possible, as follows.

2.3.3.a. ***Fixed-time labeling***

One common approach is to define a fixed time window and set a label every time that time window ends [27,63,78,79,96,99,117]. This approach is frequently adopted in affective computing experiments where the subject is monitored in an open world experiment (where the subject performs the usual tasks he or she performs during his or her everyday life) [228]. The number of registries obtained from a subject depends directly of the experiment duration.

$$\mathrm{N}^{\circ}\mathrm{registries} = \frac{Experiment\ length}{time\ window} \tag{2.1}$$

In this sense, we can see a wide variety of time window lengths, ranging from 20 seconds [27,28,239] to 20-30 minute time windows [117,159].

2.3.3.b. ***Context-driven labeling***

Other very commonly followed approach is the context-driven labeling. If focuses on proposing to take emotional labels when certain events happen (e.g. solving an exercise, watching a video, etc.) [14,42,88,95,130,184,186]. This may vary depending on the context, and may be discussed its suitability depending on the contexts. The number of registries obtained from this approach depends on the number of events to be labeled on the experiment:

$$\mathrm{N}^{\circ}\mathrm{registries} = \mathrm{Number\ of\ events\ during\ the\ experiment} \tag{2.2}$$

2.3.3.c. ***Signal-driven labeling***

Another possible approach to follow is to label every time a certain predefined event has been detected from the data sources used (e.g. heart rate has increased a 40% in less than two minutes, the subject started crying, the subject stopped using the mouse [236].Other approaches that might be considered are the subject has turned his or her face down three times while solving a given exercise, or in case of subjective labeling, the user feels that is going through a new affective state). Some works may use the amount of data collected to trigger a label request, as happens in some works using the number of keyboard interactions as a criteria to get a label [123].

As seen in the proposed examples, the predefined events usually depend also on a fixed time or an event time window to be recognized (some signal or behavior detected in a given time window) but can also be time-independent (e.g. the subject started crying, or the subject himself or herself notices an emotion change).

Depending on the phenomenon to be detected to trigger the labeling, this trigger could be automated (if thresholds are defined for measurable events) or manual (for hard-to-automatically-detect events such as crying or self-detected changes). In case of subjective triggered labeling, this approach may force the subject to be constantly self-evaluating his or her emotions, being this a disturbing factor when dealing with a task.

It is not possible to predict the number of registries that this approach will provide from each subject, as depends on the reactions detected during the experiment.

2.3.4. Time of the labeling

Another factor to decide when designing an experiment is when the labeling is to be performed. There are two different alternatives to take, during or after the experiment. Both options are also compatible, carrying out the labeling during the experiment and refining it afterwards.

2.3.4.a. *During the experiment (in vivo labeling)*

When performing the labeling during the experiment, the whole context can be taken into account.

If the labeling is being performed by the subject[43,78,113,236], in this case he or she may be more conscious of his or her current feelings than afterwards. In case an external annotator is responsible for the labeling, he or she may not be aware of all the factors involved in the experiment, that is why, in this case, the expert should be provided of all the means to be informed of all the possible factors of the experiment without disturbing it. With this live expert labeling scenario in mind, was designed the BROMP methodology was designed [162].

2.3.4.b. *After the experiment (post facto labeling)*

This approach relies on labeling an experiment reviewing it from the data recorded during the experiment [27,28,42,44,89,240] (the more data recorded, the better) allowing a detailed analysis of all the factors involved during the experiment.

In this case, some factors are important when reproducing the experiment, such as data synchronization or a tool that provides the functionality of navigating through the experiment [17](coarse and fine-grained movements through the experiment allowing to reproduce everything to the maximum detail).

2.3.5. Labeling presentation

Another important factor (especially when the responsible for the labeling is the subject himself or herself) is the way the labeling is being asked. As seen in the different emotional labeling formats in the corresponding section (2.3.1), the most common approaches are the categorical one and the dimensional one.

For the categorical one, the most common approach is showing a list with the emotions, but some factors should be taken into account such as the format (all the emotions should be written with the same font, size and background) and the order they are displayed (maybe giving more importance to those presented first).

For the dimensional approach, some standardized ways of asking for the labeling have been proposed, such as the widely adopted Self-Assessment Manikin (SAM) [39], a series of graphics that illustrates the different values of the dimensions valence, arousal and dominance.

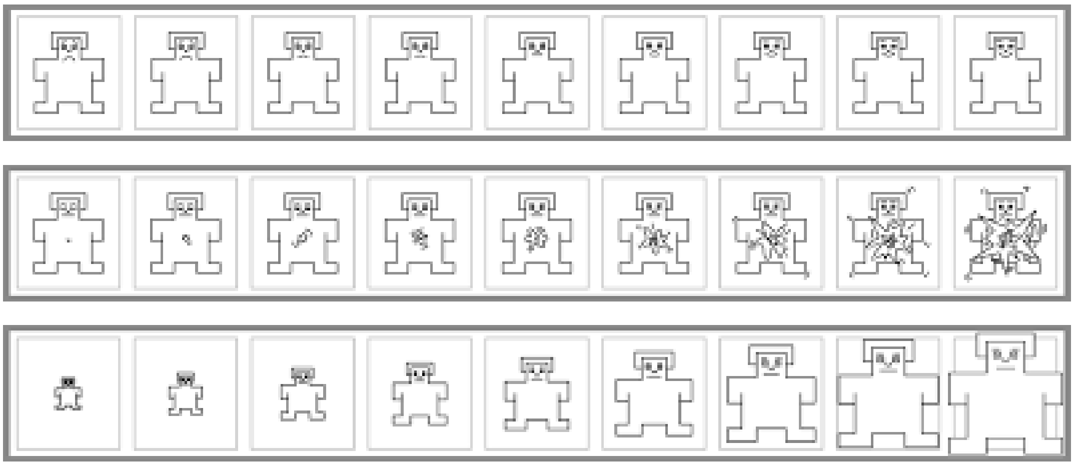


Figure 6 The Self-Assessment Manikin Scales for valence, arousal and dominance [39].

When asking for the labeling to the user, other key point to consider is if the questionnaire will be showed with some instructions about it or if the way to provide the labeling will be explained before the experiment starts. In case an explanatory text is to be shown with the form, the text to be shown should be carefully written, trying to avoid any influence on the subject’s choice. Also the number of times the explanatory text will be shown should be defined (for instance, show it just one time to avoid possible user experience problems caused by a huge text that may affect the emotions, or show a big explaining text the first time and a summary the following times).

2.4.Emotional elicitation method and task proposed

As the main goal of the experiments performed in related works is collecting data with emotional detection purposes, a common point we can find in many of them is the elicitation of some affective states during the experiments proposed. In this sense, we can group related works into different categories depending on the nature of the elicitation methods used.

First, we can find some works that aim to collect data from a regular computer use, where the experimental subjects are not told to do anything but their regular tasks [76,117,123]. These works usually aim to collect mouse and keyboard data in data collection experiments that take several days (as, in case a several-day experiment is to be carried out with physiological signals, the setup has to guarantee the synchronization of all the data sources).

Second, some works that propose an educational related tasks as emotion elicitator. These education related elicitators include using a tutoring system [14,63,96,99,126,230], carrying out tasks in a second language [51,136], essay writing [27,28], Mathematics or related tasks [113,137,148], etc. It is common to find in these works, when using a discrete set of emotions to detect, some states closely related to attention and learning such as stress [136], attention [14], anxiety [51], confusion [79], boredom [28,79], engagement [28], etc.

Third, works that use different stimuli as emotion elicitators. There is a wide variety of stimuli used in related works, such as images [53,131,184,242], audio clips[155] or music clips[49,88,186], video clips[44,227,249], or even colors [4]. Although some works use non-standardized stimuli (i.e. the proposed materials have not been validated with a population of similar characteristics to the subjects of the experiment), some of the most common standardized stimuli used in these related works include the International Affective Picture System (IAPS) [127] or the International Affective Digitized Sounds [40].

And last, there are works that use other elicitation method as can be videogames [89,171], web browsing and online shopping [100] or even programming [79].

Table 3 shows the different emotional elicitators used in related works, as well as labeling details from the works introduced in this section:

Ref.	**Year**	**Elicitation Method**	**Educational context?**	**Labeling used**	**Labeler**	**Labeling during or after the experiment**	**Data sampling temporal criteria**
[81]	2017	First experiment: Emotional labeling; Second experiment: Web search.	No	Next action to be taken by the participant	Automated (according to the next action performed)	During experiment	Each mouse action
[136]	2017	Fixed text in familiar and unfamiliar languages (with different length and familiarity	Yes	Stress perception	Participant	During experiment (after each task)	Each task

Ref.	Year	Elicitation Method	Educational context?	Labeling used	Labeler	Labeling during or after the experiment	Data sampling temporal criteria
[227]	2017	Multimedia materials (9 video clips, pictures and texts) -we presented these selected videos, images and texts to induce certain emotions and then prompted the user with some related questions. The answers to the questions were 50 to 100 characters long -There were some multiple choice questions, for which participants' used the mouse clicks also.	No	10 different emotional states: amusement, happiness, inspiration, surprise, sadness, sympathy, anger, disgust and fear and neutral 2 different emotional groups of emotions: positive valence or pleasant (amusement, happiness, inspiration and surprise) and negative valence or unpleasant (sadness, sympathy, anger, disgust and fear)	Participant	During experiment	Each task
[123]	2018	Normal computer use	No	Seven emotional states: happiness, sadness, boredom, anger, disgust, surprise and fear Biometric labeling	Participant	During experiment	Every 600 keyboard events
[184]	2017	IAPS affective images (45 images)	No	Affective valence and arousal	Participant	During experiment	Each task (each iaps image)
[14]	2017	Two e-learning courses with music	Yes	Learner's attention	Participant	During experiment	Each task
[53]	2017	IAPS affective images (15 images)	No	Affective valence, arousal and dominance	Participant	During experiment	Each picture

Ref.	Year	Elicitation Method	Educational context?	Labeling used	Labeler	Labeling during or after the experiment	Data sampling temporal criteria
[4]	2016	Color during learning	Yes	Color Emotion Scales (dark/light, pleasant/unpleasant, fresh/stale, heavy/light, calm/exciting, dull/sharp, tense/relaxed, warm/cool, interesting/ boring) and performance	Participant	During experiment	Each task
[249]	2017	Music video	No	Affective valence, arousal, dominance and liking	Participant	During experiment	Each task (music video)
[88]	2016	Music	No	Five affective states: happy, sad, scary, peaceful and other	Automated (each set of audios correspond to an emotional category)	During experiment	Each task (block of audio excerpts)
[49]	2017	Music	No	Four affective states: Happy, sad, nervous and bored	Participant	During experiment	
[125]	2014	IAPS affective images (20 images)	No	Five affective states: sadness, disgust, fear, happiness and neutral.	Participant	During experiment	Each task
[99]	2015	E-learning tasks (MetaTutor)	Yes	Eight affective states: Happy (enjoyment, hope, pride, curiosity and eureka), Anger (frustration), Neutral, Fear (anxiety), Surprise, Disgust, Contempt, Sadness (hopelessness, boredom).	Participant	During experiment	Every 14 minutes
[148]	2014	Mental arithmetic task	Yes	Cognitive stress, rest	Automated (task 1 data labeled as rest, task 2 data labeled as cognitive stress)	During experiment	Each task

Ref.	Year	Elicitation Method	Educational context ?	Labeling used	Labeler	Labeling during or after the experiment	Data sampling temporal criteria
[242]	2017	Dynamic images (48 images)	No	Affective valence and arousal	Participant	After experiment	Each task (group of images)
[51]	2017	E-learning task (English as a Second Language)	Yes	Anxiety	Participant	After experiment	Each task
[89]	2017	Videogames	No	Affective valence and arousal	Participant	After experiment	Each gaming session
[238]	2017	Login credentials data (the data comes from a dataset collected by others)	No	Fatigue	Automatic labeling (according to the trials performed)	During experiment	Each task (password trial)
[27]	2013	Essay writing	Yes	Fifteen affective states: anger, contempt, disgust, fear, happiness, sadness, surprise, boredom, confusion, delight, engagement, frustration, anxiety, curious, and finally and neutral.	Participant	After experiment	Fixed time interval (15-second interval)
[79]	2012	Programming problems	Yes	Confusion (negative valence, positive arousal), boredom (negative valence, negative arousal) and a special emotion state of "others"	Researcher	After experiment	Fixed time interval (20-second intervals)
[117]	2008	Normal computer use	No	Affective arousal and valence	Participant	During experiment	Fixed time interval (20 minutes)
[116]	2013	Music and Programming tasks (in an object-oriented language used to teach programming to children)	Yes	Affective arousal (based on galvanic skin response)	Automatic labeling (according to GSR)	After the experiment	Fixed time interval

Ref.	**Year**	**Elicitation Method**	**Educational context?**	**Labeling used**	**Labeler**	**Labeling during or after the experiment**	**Data sampling temporal criteria**
[42]	2015	Question answering	No	Cognitive task	Automatic labeling (depends on the question)	After the experiment	Each task
[28]	2013	Academic essay writing	Yes	Three affective states: boredom, engagement, or neutral	Participant	After the experiment	Fixed time interval (every 15 seconds)
[78]	2011	Normal computer use	No	15 affective states: frustrated, focused, angry, happy, overwhelmed, confident, hesitant, stressed, relaxed, excited, distracted, bored, sad, nervous and tired.	Participant	During experiment	Depending on the user's interaction activity level, taking into account the 10 minutes of interaction previous to the labelling
[122]	2015	Essay writing	No	Seven affective states: happiness, sadness, boredom, anger, disgust, surprise and fear.	Participant	During experiment	Depending on the user's interaction (every 600 events registered)
[131]	2014	Affective images (IAPS) and fixed text (numerical)	No	Affective valence and arousal	Participant	During experiment	Each task (affective image and typing text "24357980")
[137]	2014	Mental arithmetic questions with time limit	Yes	Cognitive stress	Participant	During experiment	Each task
[228]	2013	Fixed/free text	No	Seven affective states: confidence, sadness, nervousness, happiness, tiredness, hesitation and neutral	Participant	During experiment	Each task (fixed text), fixed time interval (every 15 minutes in the free text task)
[237]	2013	Facial feedback induction (holding a pen with the teeth and the lips) and fixed text (numerical)	No	positive and negative states	automatic (according to the facial feedback inducted)	During experiment	Each task (facial feedback and input text)

Ref.	Year	Elicitation Method	Educational context?	Labeling used	Labeler	Labeling during or after the experiment	Data sampling temporal criteria
[100]	2016	Three experiments: i) intelligence test designed to be unfair; ii) goal-directed task on an e-commerce website; iii) online product (laptop or car) configuration	No	i) and ii) Baseline(control group)/negative emotion; iii) SAM valence	i and ii) automatic (depending on the group of the participant) and iii) participant	During experiment	Each task
[171]	2015	Computer game designed to click on buttons	No	Two affective states: confusion and content	Participant	After the experiment	Each task (each button press)
[248]	2013	Visual perception task (judgments of similarities of geometric figures)	No	Anxiety	Participant	After the experiment	experimental session
[126]	2014	Question answering in a tutoring system	Yes	Desirability, confidence and difficulty	Participant	During experiment	Each task
[236]	2011	Computer Programming Techniques course with 7 learning objects in a tutoring system	Yes	Boredom	Participant	During experiment	Interaction(signal) -driven (after a 10, 20, 30 or 40 sec pause on moving the mouse)
[230]	2012	Multimedia training package called Tactical Combat Casualty Care in a tutoring system	Yes	Affective arousal, valence and dominance	Participant	During experiment	Each task
[240]	2013	Sounds and Stroop Test	No	stress	participant	After the experiment	Each task

Ref.	Year	Elicitation Method	Educational context ?	Labeling used	Labeler	Labeling during or after the experiment	Data sampling temporal criteria
[43]	2013	Affective images (IAPS)	No	Affective arousal and valence	Participant	During experiment	Each task
[186]	2010	Music samples	No	Free report clustered as positive or negative	Participant	During experiment	Each task
[61]	2001	Imagery scripts	No	affective arousal and valence, vividness of the imagery	Participant	During experiment	Each task
[251]	2006	Stroop test	No	Stress	Auto (depending on the task)	During experiment	Each task
[96]	2010	Two e-Learning course materials	Yes	Interactive/Non-interactive material	Auto (depending on the task)	During experiment	Fixed time interval
[44]	2010	Film fragments	No	Affective arousal and valence	Participant	After the experiment	Each task
[63]	2014	A web-based ITS for math, statistics, science, and other domains	Yes	12 affective states: anger, disgust, contempt, happiness, surprise, anxiety, confusion, boredom, curiosity, eureka, engagement/flow, frustration and neutral.	Participant	During experiment	Fixed time interval
[113]	2014	Problem-solving activity (trigonometry, backward digit span, and logic)	Yes	Four affective states: Stress, confusion, frustration and boredom	Participant	During experiment	Each task
[155]	2014	Affective sounds (IADS)	No	Five affective states: high arousal positive (HE+), high arousal negative (HE-), low arousal positive (WE+), low arousal negative (WE-) and neutral (NE)	Participant	During experiment	Anytime
[95]	2012	Affective images (60 IAPS images)	No	Affective arousal and valence	Participant	During experiment	Each task

Ref.	Year	Elicitation Method	Educational context ?	Labeling used	Labeler	Labeling during or after the experiment	Data sampling temporal criteria
[239]	2016	Activity in a programming course	Yes	Three affective states: Boredom, Confusion and frustration	Three trained labelers	after experiment	Fixed time interval (15 second)
[159]	2014	Normal computer use (free text) and fixed text typing	No	Seven affective states: anger, disgust, guilt, fear, joy, sadness and shame	Participant	After (fixed text) and during (free text) experiment	Each task (fixed text) and fixed time (30 minutes in free text)

Table 3 Elicitation methods and labeling description used in related works.

As we can see, there is a wide variety of approaches in literature. Regarding the elicitation method the most common approach is using affective images extracted from the IAPS [127], but there are many other options found in related works. The remaining variables in Table 3 have been summarized in Figure 7.

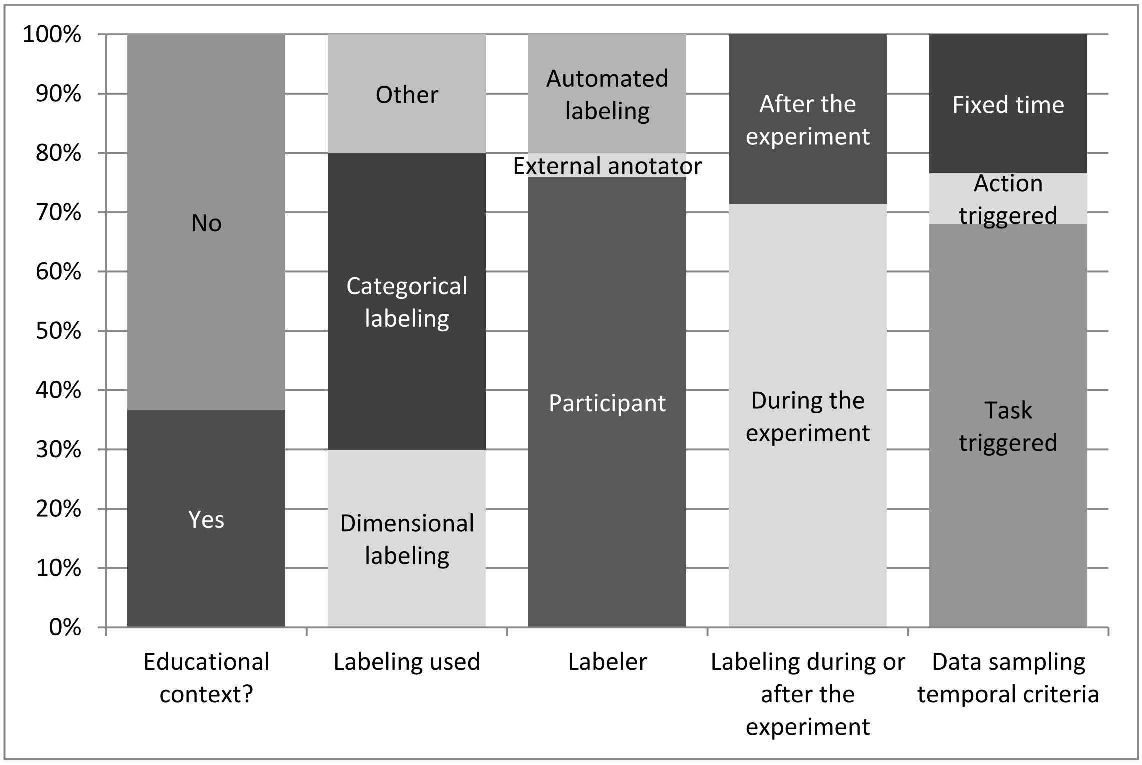


Figure 7. Labeling choices distribution on selected works.

As we can see, in the selection of works performed there are more non-educational works than educational works, as most of related works do not take place in an educational context. Regarding the labeling used, there are slightly more categorical labeling works than dimensional labeling (while there is a minority of works that aim to label any other phenomena). Most works reviewed use a labeling performed by the participant, as well as, most works also collect the labeling while the experiment is taking place (being a common approach to ask regularly the participant to provide the labeling during the experiment). Regarding the time window covered by the labeling, most approaches are based on tasks or materials, being the time spent with that task or material the time labeled. There is a small amount of works (most of them those relying on mouse or keyboard) that perform a labeling that covers a given number of actions triggered (or data instances). There is also a small amount works that use a fixed time window for the labeling, (with different fixed time window lengths).

As happened with the data sources to be used, most works adopt an approach and evaluate it, but there is a clear lack of comparison between labeling approaches. That is why this work will research in that area, using and comparing different labeling approaches in order to evaluate the impact those different methodological choices in labeling might impact the results from the models generated.

2.5.Related states of the art

As the affective state detection has been a research topic during the last years, some works have compiled the information from many related research. In this sense, we can find several works that analyze different subsets of related works based on certain characteristics.

We can find some analysis of the state of the art focused on the data sources used to perform affective states detection. In [120] a series of 11 works that aim to carry out affective state detection by means of analyzing keyboard and mouse interactions is analyzed. Other common data sources in affecttive computing works such as facial expressions is used as criteria to analyze the works included in [144,154]. During last years, many works have faced the problem of affective state recognition using brain-computer interaction devices or EEG devices as can be seen in [3,157]. A collection of works using body movement as source of affective information is evaluated in [185]. The applicability of affective computing to mobile environments is being also widely researched nowadays, as can be seen in [177,252].

Other surveys on affective state detection focus on the domain to be applied in order to filter the works to be evaluated. As aforementioned, the educational field has a special interest in affective computing and we can find an evaluation of different works aiming to detect emotions in educational contexts in [205,246].

We can find also surveys that use other filtering criteria to choose the works to be evaluated. While [90] focuses only on different approaches and technologies for stress detection, [180] focuses on the different analysis performed to the data with special attention to the fusion methods used in multimodal approaches.

2.6.State of the art conclusions

As we have seen in these related works, there is a wide field of research in the affective state detection. We have focused on a limited set of works, i.e., those works using keyboard, mouse, or some physiological signals. Although we have not limited our search to research works applied to the educational domain, we have set the focus on how those works perform the affective state detection, and how some issues, such as the emotions elicited, can be influenced by the domain of application. A brief summary of the methodological points addressed in this section can be found in Table 4.

Elicitation Method		Emotional Labeling	[81]
		Web Browsing.	[81,100]
		Fixed Text Writing	[131,159,228]
		Video	[44,227,249]
		Pictures	[43,53,95,125,131,184,227,242]
		Text Reading	[227]
		Normal Computer Use	[78,117,123,159]
		E-Learning Related Task	[4,14,27,28,42,51,63,79,96,99,100,113,116,122,126,136,137,148,228,230,236,239]
		Music And Sounds	[14,49,88,116,155,186,240]
		Color	[4]
		Computer Games	[89,171]
		Facial Feedback Induction	[237]
		Visual Perception Task	[248]
		Stroop Test	[240,251],
Labeling Format	Affective Dimensions	Valence	[43,44,53,61,89,95,100,117,131,184,227,230,242,249]
		Arousal	[43,44,53,61,89,95,116,117,131,184,230,242,249]
		Dominance	[53,230,249]
		Groups Of Valence + Arousal Scores	[79]
	Affective Categories	Amusement	[227]
		Anger	[27,63,78,99,122,123,159,227]
		Anxiety	[27,51,63,99,248]
		Boredom	[27,28,49,63,78,79,99,113,122,123,236,239]
		Cognitive Stress	[42,137,148]
		Confident	[27,63,78,79,113,126,171,228,239]
		Contempt	[27,63,99,100,171]
		Confusion	[27,63,79,113,171,239]
		Curiosity	[27,63,99]

	Delight	[27]
	Desirability	[126]
	Difficulty	[126,227]
	Disgust	[27,63,99,122,123,125,159,227]
	Distracted	[78]
	Engagement	[27,28,63]
	Enjoyment	[99]
	Eureka	[63,99]
	Excited	[4,78]
	Fatigue	[238]
	Fear	[27,99,122,123,125,159,227]
	Focused	[78]
	Frustration	[27,63,78,99,113,239]
	Guilt	[159]
	Happiness	[27,49,63,78,88,99,122,125,227,228]
	Hesitant	[78,228]
	Hope	[99]
	Inspiration	[227]
	Joy	[99,159]
	Nervous	[49,78,228]
	Neutral	[27,28,63,99,125,155,227,228]
	Other	[79,88]
	Overwhelmed	[78]
	Peaceful	[88]
	Pride	[99]
	Relaxed	[4,78]
	Rest	[4,148]
	Sadness	[27,49,78,88,99,122,123,125,159,227,228]
	Scary	[88]
	Shame	[159]
	Stressed	[78,113,136,137,148,240,251]
	Surprise	[27,63,99,122,123,227]
	Sympathy	[227]
	Tired	[78,228]
Other	Next Action To Be Performed	[81]
	Learner’s Attention	[14]
	Color Emotion Scales	[4]
	Liking	[249]

		Interactive/Non-Interactive Material	[96]
Label Source	Participant		[4,14,27,28,43,44,49,51,53,61,63,78,89,95,99,100,113,117,122,123,125,126,131,136,137,155,159,171,184,186,227,228,230,236,240,242,248,249]
	External Labeler		[79,239]
	Elicitation Method Label		[42,81,88,96,100,148,237,238,251]
Time Of The Labeling	During The Experiment		[4,14,43,49,53,61,63,78,81,88,95,96,99,100,113,117,122,123,125,126,131,136,137,148,155,184,186,227,228,230,236–238,249,251]
	After The Experiment		[27,28,42,44,51,79,89,116,159,171,239,240,242,248]
Data Sampling Temporal Criteria	Each Task		[4,14,42–44,51,53,61,88,89,95,100,113,125,126,131,136,137,148,159,171,184,186,227,228,230,237,238,240,242,248,249,251]
	Fixed Time		[27,28,63,79,96,99,116,117,159,228,239]
	Interaction Triggered		[78,81,122,123,236]
Data Source	Mouse Related Features		[81,100,116,117,126,136,137,171,227,230,236,239,248]
	Keyboard Related Features		[27,28,42,78,79,116,117,122,123,131,136,137,159,227,228,237–239]
	Electrocardiography Related Features		[4,14,43,49,51,53,88,89,95,125,148,184,186,240,242]
	Skin Conductance Related Features		[44,49,63,88,89,96,99,113,116,125,155,184,240,242,249,251]
	Skin Temperature Related Features		[49,125,249,251]
	Breath Related Features		[61,89,125,148,249]
	Electroencephalography Related Features		[95,113,155,249]
	Electromyography Related Features		[44,89,249]
	Electrooculography Related Features		[249]
	Blood Volume Pressure Related Features		[113,249,251]
	Facial Expressions Related Features		[99]
	Speech Related Features		[43]
	Pupil Diameter Related Features		[251]
Machine Learning Algorithm	Clustering	K-Means	[227]
	Instance-Based	K-Nearest Neighbors (K-NN)	[44,96,122,125,227,238,249]
		IB1	[227]
		Kstar	[227]
	Ensemble	Bagging	[227]

	Random Committee	[227]
	Random Forest	[27,89,125,171,227,239,239,248]
	Adaboost	[42]
Bayesian	Naive Bayes	[27,42,113,122,125,148,249,251]
	Bayesian Networks	[27,113,122]
Support Vector Machines	Support-Vector Machines (SVM)	[27,42,44,49,96,113,125,148,155,159,171,184,248,251]
	Least Square SVM (LSSVM)	[249]
Neural Networks	Neural Network	[44,122,125,159]
	Deep Learning Techniques	[249]
Regression	Logistic Regression	[117,159,171,249]
Rule System	RIPPER	[125]
	OneR	[27,76,125,186]
Decision Tree	Decision Tree Classifiers	[42,113,122,122,239,251]
	C4.5 (Or J48) Tree	[27,79,125,159,171,236,239,240]
	Reptree	[27]
	Random Tree Classifier	[27,79,159,227]

Table 4. Summary of methodological issues addressed in this work

One of the things that are most clear is the high number of methodological variables that can be identified in these works (described in the previous subsections), including: i) emotion representation method, ii) data labeling (frequency, labeler, format, etc), iii) task proposed, iv) laboratory or real world conditions, v) elicitation method, vi) data sources, vii) features generated, viii) data analysis performed, etc. It can be considered that most of those issues remain open research issues, as we can find a wide variety of approaches with very different methodological proposals (due to the lack of a clear methodology tested, evaluated and compared). We also can see the high variance of abstraction when discussing depending on what issues. Some issues can be clearly described in some works as the elicitation methods, emotion representation methods, etc. while some of them, such as the preprocessing performed on the data between the data collection and the predictive model generation, are rarely addressed.

After identifying all these methodological differences between works, it is unclear which choices should be taken due to the lack of comparison between them. Is that lack of comparison between different methodological points one of the motivations to carry out this research, in order to evaluate the impact of some of those proposed methodological choices in a affective state detection system.

This work aims to advance in the evaluation of those methodological points found in related works. The base of this work is to propose initially a methodology. The main contribution of this work, in contrast with the works found in literature, is to provide, starting from that methodology proposed, a wide evaluation of different possible alternatives in some of the methodological open points identified in related literature.

The main advance in contrast with the works found in literature is to provide a wide evaluation of different alternatives in some of the methodological points identified in literature. The methodological points identified in this literature review to be addressed in the different stages of this work include: labeling approaches (i.e. different labelers and different ways to process affective labels), different data sources combinations, different algorithms to be used for the model generation, different ways to normalize the data, different ways to preprocess the data (i.e. class distribution balancing and feature selection technique), different approaches for data processing (i.e. 2-step classification approach and clustering based feature generation). For each one of the possible configurations proposed for each one of the aforementioned methodological issues, several models are to be generated, and thus the prediction results are to be compared to evaluate the impact of that methodological variable in the affective state prediction.

3. Approach

The first step of this research was to carry out an exploratory analysis on the different ways researchers have faced the problem of affective state detection by means of machine learning techniques using a combination of different data sources. Once we have seen the different approaches for emotions detection adopted in the literature, an initial proposal of the methodology has to be made. Taking into account, initially, the open issues identified in the state of the art review, and in following experiments, those open issues found in previous experiments, we propose a multimodal approach based on processing in a combined manner different data sources that are available in an educational domain. Figure 8 depicts the work here presented (with a green background) framed in the MAMIPEC overall approach. The blue boxes represent the additions required in order to provide a traditional e-learning platform (non-blue boxes) affective related capabilities. It can be seen how, using information obtained from bio-feedback devices and the interaction with the user device, a data mining process can be performed in order to compute affective information that can be stored in the learner model. In this way, the result of that process should reflect the learner affective state, which has to be included in the learner model (from now on, affective learner model) so affective triggered feedback can be performed by the e-learning platform.

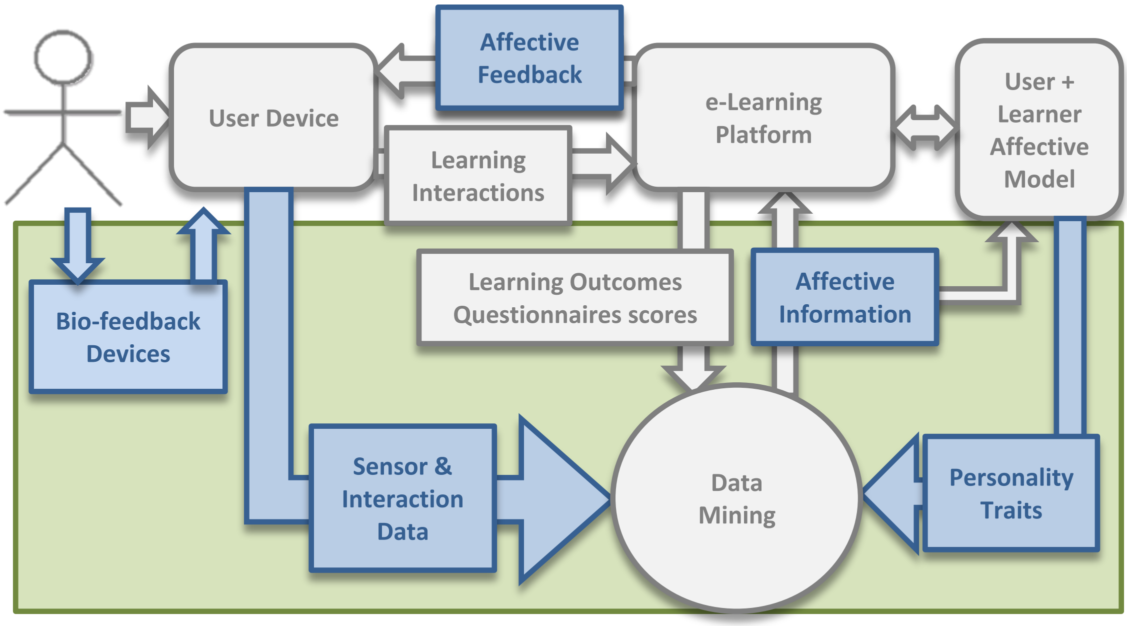


Figure 8. Proposed approach in the MAMIPEC project where this work is framed. Blue boxes depict the design of the introduction of affective information into a traditional e-learning platform. Green background represents the work to be carried out in this thesis.

With the MAMIPEC approach in mind, it has to be reminded (as already discussed in the introduction) that this Thesis focuses on the detection side and how to perform it by means of supervised learning techniques (as done in most related works depicted in section 2.2). Nevertheless, there are a significant number of methodological issues involved in the design of the system proposed in this thesis. Figure 9 depicts these issues found, with some of them being addressed along this research work.

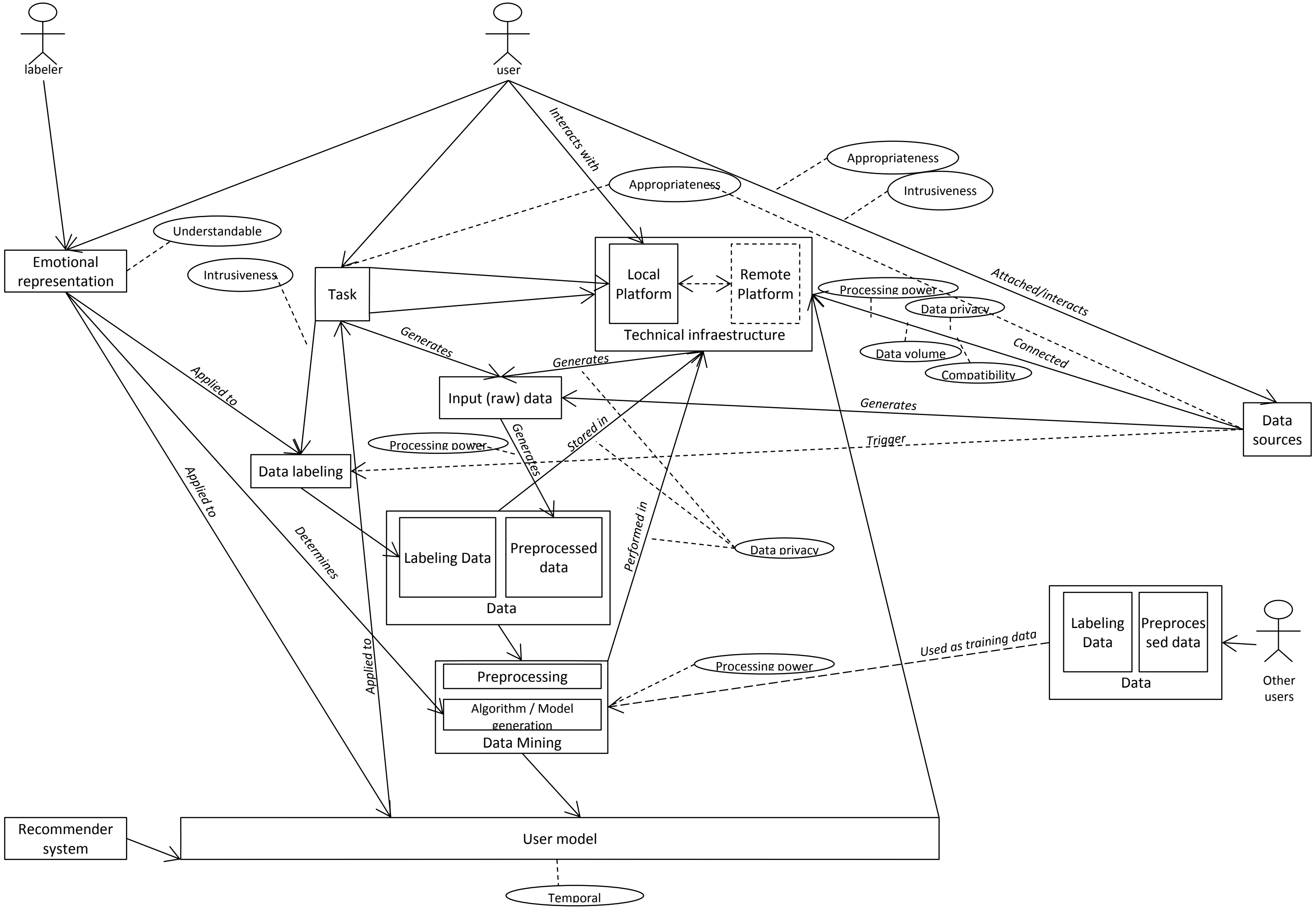


Figure 9 Methodological issues found in the development of an affective state detection system.

As we can see in Figure 9, there are many methodological issues that should be taken into account when designing an affective computing experiment. We can see the issues related to the data sources (on the right part of the figure), evaluating the appropriateness of the data source, the intrusiveness, the data processing required and where that data processing can be performed and stored (raising technological infrastructure, security and ethical issues). The importance of the task being performed during the detection as well as its relation to the technical infrastructure, the data generated and the data labeling are also issues to face when designing an experiment. The labeling itself also represents one of the open points in this field of research, raising many related concerns, such as the emotion representation chosen to model the affective state of the participants, who should label the data or the temporal validity of the label provided to the model. Related to the data itself, the source, how it is stored, preprocessed and processed should be taken into account. All these issues have been identified in the literature and depicted for the AMO-ML methodology here proposed.

3.1. Research structure

This research has been designed following an incremental approach. The overall goal is to design, build and evaluate an experimental methodology to detect affective states by means of machine learning techniques following a multimodal approach. This research has been structured into two main stages, with a small transitional stage between them. While the first stage aims to build the methodology, starting from the related literature, the second main stage aims to develop that methodology in a real-world learning scenario. During the transition stage, the first contact with the real-world learning scenario is performed. During all the stages of this work, an experiment is carried out, in each stage, evaluating different methodological open issues found. Due to the nature of the incremental approach, the evaluated methodological issues get narrower with the different stages, starting with some big open issues in the first stage and advancing in finer-grained issues related to the data analysis in the second stage. According to these issues found in this stage, different hypothesis have been stated (as seen in section 1.4).

3.1.1. Stage 1

The first stage of the research has been designed in order to perform an initial exploratory analysis in order to build the a methodological approach to detect affective states in computer users by means of machine learning techniques, following the schema depicted in Figure 8. An initial research on related literature is the first step in this stage. Once that an initial view of the field has been obtained, the methodology has to be designed, and will be defined in its instantiation in an experiment (that will be presented in section 4). Due to the exploratory proposal followed in this first stage, some methodological points that can be clearly identified in the review of related literature (spotted in section 2.6) have been researched, with two main methodological variables evaluated in this stage:

- Data sources: evaluate which combination of the proposed data sources performs better when predicting affective states (evaluating this way the hypothesis **H1** and objective **O1.1** introduced in section 1.4).
- Data labeling: evaluate which affective data labeling source provides better accuracy rates (evaluating this way the objective **O1.2** introduced in section 1.4).

Nevertheless, there are other secondary methodological aspects to take into account in this stage:

- Emotion representation: as the data to be collected has to be labeled with the emotions to be predicted
- Task proposed: as an educational task has to be performed in the experiment
- Data analysis technique used: as seen in section 2.2, data mining techniques are the resource used by those works aiming to perform detection, but there are many different algorithms used in related works.

3.1.2. Transition stage

After the celebration of the first experiment in stage 1, with an initial version of the methodological approach defined, a small experiment carried out in collaboration with University of Valencia in the frame of the MAMIPEC project was used in order to perform a first approach to define a reference experimental scenario based in real-world learning conditions. Although due to the nature of the experiment, the methodology followed in stage 1 was not fully applied in this experiment (that is why this stage is considered a transition stage). The conclusions and lessons learnt from this transitional stage helped both to propose some new methodological approaches in detection (i.e. a 2-stage based prediction approach) and to stablish a first contact with a real-educational learning scenario setup.

3.1.3. Stage 2

After the definition of a methodological approach developed in the experiment carried out in stage 1, and a first contact with a real-world learning scenario in the transition stage, a final stage is aimed to hold an experiment built from the outcomes of the previous stages. This stage has been designed in order to evaluate the applicability of the approach proposed in stage 1 in a real world scenario similar to the one described in the transition stage. In contrast to the variables evaluated during stage 1, the methodological variables to be evaluated in this stage are finer-grained, are more related to the data preparation, as a consequence of the lack of detail identified in this sense in section 2.2. In this case, the focus has been set to the evaluation of the following points:

- Interaction data normalization: The applicability of an initial interaction baseline to evaluate the potential accuracy increase of the prediction from interaction data sources. This idea is based on the application of the initial baseline used in the physiological signals in stage 1 to the interaction data sources.

- Dimensionality reduction techniques :The impact on the accuracy of the models generated of using different dimensionality reduction techniques (i.e. Forward Feature selection and Principal Component Analysis)
- Class balancing techniques: The impact on the accuracy of the models generated of using different class balancing techniques (i.e. SMOTE and Equal size sampling)
- Class attribute discretization: the impact on the accuracy of the models generated of using different data discretization approaches on the class attribute.
- Model generation variations: In this sense, the application of unsupervised learning techniques and a 2-stage modeling approach (introduced in the transition stage) have been proposed and evaluated.

3.2. **Experimental guidelines**

Each one of the iterations planned of this research includes designing and holding an experiment in order to validate the proposed hypothesis of each stage. The experiments are going to follow a common approach, as this work is framed in the affective state detection in educational context area, with some differences in order to focus on different points depending on the stage of the research.

The main goal of the experiments held in this work is the collection of data to evaluate the different hypotheses and objectives introduced in section 1.4. By collecting data, we aim to create an emotional dataset to train supervised learning techniques based models in order to predict the affective states that appeared during the experiment. To do that, the emotion elicitation plays a key role in the experimental design (so we can record the data generated from the different data sources while the participant experiences the emotional changes we aim to predict). Usually, short experiments are designed to assure the presence of emotions during the time emotions are elicited (so this way researchers also know the kind of emotions expected). Commonly used methods for emotion elicitation are based on passive events such as video or image viewing and sound listening, covering, this way, a wide variety of emotions (as seen in section 2.4). As we are dealing with educational contexts, the most common approach is making the subject to carry out a certain task in an e-learning platform [63,236], and that is going to be done in the experiments to be described here. Additionally, we also know, from related works, which are the most common emotions registered when carrying out experiments in learning contexts: boredom, confusion, frustration and engagement [178].

One of the key points in our approach is to try to get as close as possible to a real-world scenario. Due to that, the elicitation method to propose in our experiments aims to replicate possible emotion elicitators to be found in real educational tasks. In our case, the main elicitator to be used is the modification of the difficulty level of the educational tasks proposed. Other elicitators as time limits may also be used.

The structure followed in the experiments proposed is quite similar in the coarse-grained level, following the same structure in the experiments of both stage 1 and stage

2. Nevertheless, the approach and objectives in both stages differ, while the first stage aims to be an exploratory initial experiment in order to develop an initial version of the methodology, the second stage aims to identify the adaptations needed to use that methodology in a real-world learning scenario, generating, with those adaptations (as well as other methodological issues found in the previous stages), a new version the methodology proposed in this work. The common framework for both stages has the following structure:

- Experiment design:

  An initial phase where the objectives are to be matched with the design of an experiment.

  - Data sources to be used:

    This decisions aims to define which data sources should be used in the data collection experiment. This issue might have an impact on other experimental design steps, such as the task to be performed (as the data sources should allow the participants to perform the task in a naturalistic way) or the data processing step (as different data sources provide different kind of data, with a different frequency, etc.).

  - Labeling to be used:

    This issue has to be clarified to see if the data labeling has to be performed during or after the experiment. Other issue to clarify is the source of the labeling, as it could mean requiring external experts during or after the experimentation.

  - Task proposed:

    As the main goal of the experiments is collecting affective data, the task proposed should provide the tools needed to elicit different affective states. In our case, the context of the task is limited to the educational domain.

  - Elicitation method:

    One of the key points in the experimental design. Depending on the desired affective states, the elicitation method should be designed wisely. In our case, as our research is framed in the educational domain, educational-related emotions are to be elicited by using elicitation methods that can be found in the proposed domain (in our case, we chose time limits and task difficulty, as these elicitation methods can be easily found in education).

- Experiment
  - Tools implementation:

    All the tools required for the experiment to be hold should be ready. An experiment requires both a participant performing a task and the data being collected and recorded and those two processes require tools. Some tools might be available and fit the experimental constraints (in our case, some tools such as screen recorders were acquired) but other tools might be required to be developed (in our case, the transparent keyboard and mouse logger was developed as well as the tool to synchronize the data recorded from different data sources).

- Experimental set setup:

  As in our experiments the presence of the subjects was required, different computers had to be set up in order to have all the systems ready to perform the data collection while the participant performed the proposed tasks.

- Data sources placement:

  As aforementioned, there is a relation between the task proposed and the data sources used. Some data sources might have a negative impact on the task solving process. The data source to be used is not the only methodological variable that might be discussed, as a single data source can be placed in different parts of the participants, having the placement of the sensor an impact both on the performance of the task as well as on the quality of the signals recorded [67].

- Baseline:

  Due to the variance on what normal physiological values can mean on every different person, to analyze physiological data, it is commonly recorded a baseline in order to see the values of a person in a “neutral state”. One of the main contributions of this work is transferring this approach from the physiological signals to the interaction devices.

- Task solving:

  The design of the task order has to be designed, taking into account that aspects such as the emotion elicitation can drive that order.

- Data analysis
  - Data cleaning:

    After the experiment has finished, the data collected might need to be cleaned. Some data sources can induce noisy data that might have a negative impact on the models generated. The data cleaning has to be performed taking into account the nature of the data, so the noise of each data source should be cleaned in a personalized way to each data source.

  - Feature generation:

    In order to generate the models, some features have to be generated from the raw data recorded. Each data source requires a different way to generate the features due to the different phenomena recorded.

  - Data preprocessing:

Once the dataset with the features generated from the raw data is available, some preprocessing might be needed. It is common to deal with some problems such as high dimensionality or redundant features that should be faced in this stage of the data analysis.

- Model generation:

With the data ready to be processed, it is used as an input with different data mining algorithms. The models are going to be generated with 10-fold cross validation in order to avoid the overfitting of the models.

- Model evaluation:

The prediction results of the models are going to be compared and evaluated in order to see which models performed better (and what methodological variables impacted on those results).

The details and differences found in the instantiation of the steps here proposed in the different stages are to be discussed in the following sections.

4. Stage 1: Methodological Exploratory Analysis

This first stage of the research aims to set the base of this work. An initial exploratory analysis is carried out based on the works identified in section 2, and from that, a methodological approach to perform affective state detection is to be done. This methodology is set to be presented in this section, describing all the steps followed to carry out an experiment with several participants. Additionally, the first open issues found in some aspects of the methodology (e.g. data sources, data labelers, etc.) are going to be set as variables of research.

4.1. Goals

As discussed in the previous sections, the goal of the work carried out in this stage is to define an initial version of the AMO-ML methodology by means of carrying out an experiment to perform affective state detection. In addition, a series of open issues found during this methodological design following the guidelines found in section 2 are going to be evaluated.

This stage will be focused on the methodological investigation and development of a first approach to the detection of affective states of learners in online educational contexts. The first step in this stage has been, then, carrying out an analysis on how related works approached the problem of affective state detection. From that analysis (which final outcome has been the content in section 2), an initial version of the approach to be proposed has been set, as well as the identification of some crucial methodological points identified from the review of the state of the art performed. Due to the lack of prior contact with the field, only some of those more visible open research topics in the area are going to be evaluated.

To get there, an initial experiment is going to be carried out in an educational context to get affective data, so an educational-related elicitator to cause emotions on the learners has to be chosen. As the system is going to perform the affective state detection by means of supervised machine learning techniques, it is important to design the data to collect and to use as label attribute. In order to do that, and after a review of the state of the art, a selection of some of the different data sources found in literature is to be used in order to evaluate the best combination to perform our affective state detection (**O1.1**). In a similar way, different approaches in data labeling are to be evaluated in order to see which data labeling source provides better affective predictions (**O1.2**). Finally, we would aim to evaluate the first hypothesis (**H1**) proposed in section 1.4 by analyzing the results on using supervised data mining techniques on multimodal data sources in order to improve the accuracy when detecting affective features in educational contexts in comparison with single data sources.

Data mining techniques are to be used for the model generation, so a huge dataset needs to be created, which implies that a large number of participants is needed. For this reason, we should set out tasks to be solvable by as many people as possible regardless their cultural background.

This section describes the context in which the experiment was framed, its design, the participants involved and the actions done on the data gathered, in particular, how they were recorded, prepared and processed. The next section presents the results of the analysis. A discussion of the findings comes afterwards.

4.2. Methodological variables

During the exploration of related works and the design of the experiment, a series of open issues have been identified. One of the main goals of this stage is to perform an initial exploration on those different methodological issues identified and study the different approaches that can be followed in those issues. In this section, those open issues that are going to be researched in this first stage are going to be introduced and described.

4.2.1. Data Sources

The data sources considered in the proposed multimodal data mining approach are those that have been widely reported in the literature. Each of them is described next, as well as discussed the way to gather it. The indicators here commented that can be obtained from each of them are of relevance for the emotions detection.

4.2.1.a. *Keyboard*

As shown in section 2.1.2.a, the influence of emotions on the way we type can throw some information about how we are feeling in each moment [28,142,237,241]. In addition, it is a cheap and unobtrusive way of collecting data from the user.

4.2.1.a.i. Data description

From this data source, what we expect to get is all the interactions the user performs with the keyboard. Due to the nature of the keyboard, the information we want to process comes from two different interaction events:

- Key press: when the user pushes a key.
- Key release: when the user stops pushing the key, releasing it and going this to its normal position.

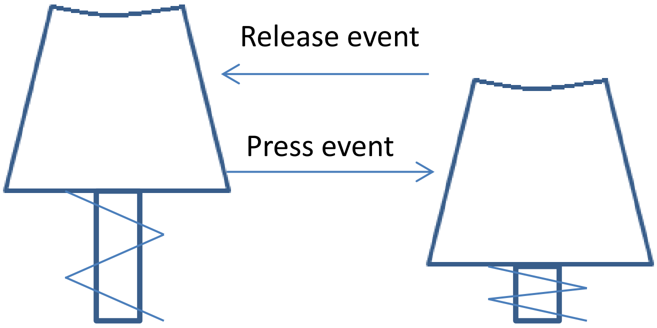


Figure 10. Keyboard interaction events

A keystroke registered will be considered the consecution of a key press and a key release of the same key. It also should be noticed that a key press and the release of that key do not have to be two consecutive events, as there can be some events between them, i.e., when a user presses the shift key for a while in order to type some uppercase characters.

4.2.1.a.ii. Discussion on gathering keyboard data

There are many important factors that should be taken into account when using key interactions as a data source.

To start, the skill level when typing of every user should be known [28] when processing this kind of information. Thus, a base line to know the typing skills of each user in the keyboard being used (as this is a factor that also could affect the data obtained) is recommended (also to evaluate the changes in typing behavior) [241].

In addition, the nature of the task and the keyboard usage needed during that task, as some task can be solved only using the mouse or another input methods. In case various input methods can be used, it should be also evaluated. That is why a multimodal approach is richer in this case than using only a keyboard as input, providing a better generalization of the generated models [78].

Another feature that could affect the data gathered by means of keyboard is not only the nature of the task itself, but also, if the learner uses the keyboard for other tasks at the same time. For instance, people visually impaired usually uses the keyboard not only with a text input purpose but also with a navigating purpose instead of using the mouse [211].

Finally, due to the nature of the tool used, although being unobtrusive from the interaction point of view, it is a very obtrusive data source from the privacy point of view. Thus, in case of using this data source in a natural context, we should make sure that the user is aware of the information we are getting from him or her and offer a ethically responsible processing for that information, looking forward storing the less explicit information possible from the information the user is typing (e.g., when the user is typing a password, the mere fact of storing that the user has pressed only key numbers is something that could be obtrusive).

4.2.1.b. *Mouse*

Another interaction tool commonly used when interacting with a computer is the mouse. Although there are fewer studies of its usage as tool to give information about emotions than studies using keyboard, there are some works in that direction [19,114,115].

4.2.1.b.i. Data description

It has some similarities with keyboard as is a commonly adopted interaction device, being cheap and unobtrusive [115], but when analyzing the way the user interacts with it (and then, the indicators that can be extracted from those interactions), its interaction behavior is quite different. We could say that there are three different ways of interacting with a mouse:

- moving it
- clicking its buttons
- moving its scroll

These different kinds of interactions are identified assuming the most common mouse design nowadays, being this a mouse with a scroll and two buttons. Based on this, the indicators to be extracted from those interactions should model them as close as possible.

For the first way of interacting (see Figure 11 for details), based on the mouse movement, it can (and is commonly) logged saving the mouse cursor position and the timestamp of that position [236]. This positions can be stored with a certain regularity (given a predefined frequency and logging the position in equal time periods) [236] or always a movement has been detected (logging the position of the mouse if this has changed more than a given distance, for instance, 1 pixel) or an hybrid approach, capturing the mouse cursor position with a given frequency if a movement has been detected. Depending on the mouse movements registered, the first approach can need more disk space resources in case the mouse is rarely moved, as its position will be stored anyway, or less disk space in case the mouse is being moved very fast (so the distance threshold stated in the second approach can be covered several times in the period stated in the first approach). Anyway, the space required to do this logging is very small given the hard disks currently available as every position log can be a plain text line. So for logging the mouse cursor position, another thing to be discussed is the units to be used to log the position. The most common unit used is the pixel, but it could also be used a common distance unit as centimeters or inches, or a normalized unit as the percentage of the total distance of the computer screen allows the mouse cursor to be moved. It should be taken into account that some factors also should be known for each dataset used, as can be the resolution of the monitor used by the learner and even the number of monitors.

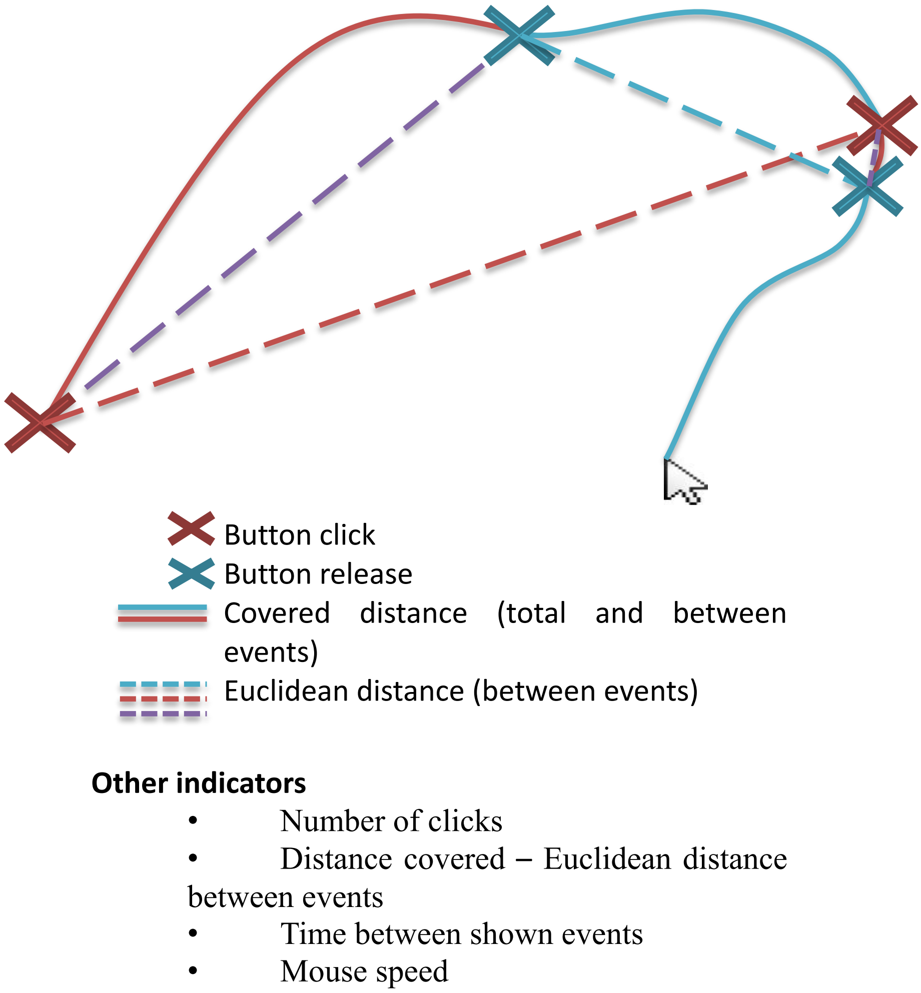

Button click
Button release
Covered distance (total and between events)
Euclidean distance (between events)
Other indicators
• Number of clicks
• Distance covered – Euclidean distance between events
• Time between shown events
• Mouse speed

Figure 11. Indicators to be extracted from mouse movement.

When logging the second way identified of interacting with a mouse (i.e., clicking its buttons), an approach similar to the one proposed in the keyboard logging section can be used, but in this case, not only should be logged the button clicked identifier with its corresponding timestamp, but also the position of the cursor when the button is clicked should be stored. A mouse click will considered the consecution of a button press and button release event, being saved these two events for every button. By logging these two events, we could also know when the user has interacted with a drag and drop element or selected text, being necessary to distinguish in these two different interactions some context information (although this could be inferred in many cases from the following interaction performed by the user, being the text selection usually followed by actions such as copying, cutting or deleting, but as these can be performed with keyboard or mouse, is quite laborious to develop a system able to do that).

To log the third way identified of interacting with a mouse (i.e. scrolling), not only the timestamp of the event should be stored, but also, the position of the event and the direction of the scroll performed. In case the mouse being used has different scrolls (e.g. horizontal and vertical scroll), the scroll the interaction that has been performed should also been identified.

4.2.1.b.ii. Discussion on gathering mouse data

As happens when capturing keyboard interactions, there are also some issues to deal with when capturing mouse interactions, some of them are common.

The first issue to deal with is, as happened with keyboard, the evaluation of the behavioral changes when interacting with a mouse by means of comparing the values of those interactions with a user base line in order to detect behavioral changes from that base line.

The nature of the task being handled is also a factor that should be taken into account, knowing the mouse usage required for solving that task and if there are alternatives to solve that task using the mouse. This factor depends on the way the application being used has been designed and implemented, as well as the way the elements to be interacted using the mouse have been disposed on screen (distances between buttons, moving elements, if there is a scroll bar or the user needs to use the mouse scroll to view the whole content, or even if the content has been displayed in a single long screen or in different pages, etc.).

In addition, this logging method is unobtrusive from an interaction point of view. However, the mere fact of knowing that all your mouse interactions are being recorded can be “psychologically” obtrusive and from a privacy point of view, arguable, so the way these interactions are recorded and stored is something that should be done taking care of these elements.

Finally, although in this work it has been proposed the use of a traditional mouse (with buttons and scroll), nowadays, there are many different ways to handle a cursor in a device, such as trackpad or touchscreens (widely used), or new ways of interaction as air gestures via devices such as Kinect or Leap motion, making the device used change completely the values reflecting the behavior of the user (this values can even change depending on the mouse model used). This is a very important point as many people with some disabilities may not be able to use a traditional mouse (for instance visually impaired people or people with mobility problems) but use any other alternative (keyboard, Kinect, etc.).

4.2.1.c. *Physiological signals*

One of the most common data sources seen in literature in the last years is the use of physiological signals obtained through bio-feedback devices to identify affective states. Although the devices needed to get this information used to be very expensive, in the last years a lot of different ways to measure some physiological signals in a no so expensive way have appeared. This change has made new technological movements emerge, such as the quantified-self one, encouraging the creation of wearable devices able to measure some activity or signals from our body [232]. It should also be mentioned some open hardware alternatives that also appeared in the last years, bringing people the possibility to create their own low-cost physiological devices [206]. It also should be said that this is possible depending on the signal to be recorded, as there are many different signals, each one with different characteristics. The ones considered in our approach are described next. After that, we present the data gathering tool used to collect these signals.

4.2.1.c.i. Heart rate

The information that can be collected from the heart rate and the issues to be considered in the process is described next.

4.2.1.c.i.1. Data description

The heart rate (HR) signal reflects beats per minute (bpm) of a person's heart and is recalculated on each beat detected on the electrocardiogram (ECG). Another common indicator extracted from the HR is the Inter-Beat Interval (IBI), which indicates the time interval (commonly expressed in milliseconds) between two following beats (so its value varies from beat to beat. This measure is also known as the RR interval, as represents the time between two following R waves (see Figure 12).

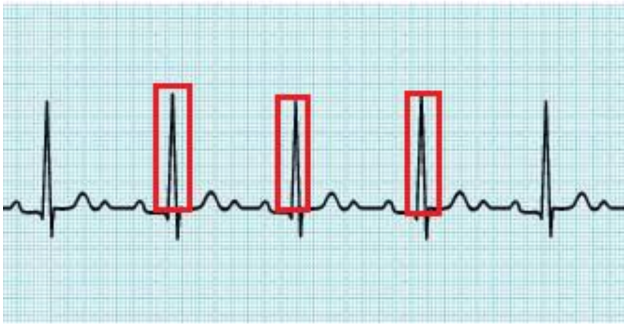


Figure 12. An electrocardiogram. R waves are contained in red squares.

Other indicator to extract from the IBI is the Heart Rate Variability, which measures the IBI variability. This way, we can measure the circulatory system activity, which is a part of the autonomic nervous system, containing a) the parasympathetic nervous system, responsible for causing a relax and calm state not only after basic functions (as can be digestion or sexual intercourse) but also after a state of tension, and b) the sympathetic nervous system, which increases the heart rate as a part of the reaction in a fight-or-flight situation (when our body feels that has to be alert).

4.2.1.c.i.2. Discussion on gathering heart rate data

When recording the heart rate, there are many factors that may affect this signal. In particular, two different signal variations are to be avoided.

On the one hand, noise, signal variations due to the sensor and its use in a non-properly way, variations that can be generated in many ways, from moving the cables of the device (moving the feet while being sensed) to having electronic devices close that may induce some noise on the signal.

On the other hand, and due to the nature of the signal, variations in the heart rate caused not only by affective state changes but by other causes as having just ended practiced intense physical activity, measuring heart rate during the digestion after a heavy meal, being in a context with some temperature changes, etc.

4.2.1.c.ii. Skin conductance

The information that can be collected from the skin conductance and the issues to be considered in the process is described next.

4.2.1.c.ii.1. Data description

The skin conductance or galvanic skin response (GSR) is a method of measuring sweating changes in skin reflected on its electrical conductance. Skin conductance reflects activity of the sympathetic nervous system, responsible of the physiological reactions to situations like stress or excitation.

As is a continuous signal, and, unlike ECG, the GSR has no patterns. For this reason, the indicators extracted from the GSR usually reflect its behavior during a given time period, as can be the mean value, the range covered, etc.

4.2.1.c.ii.2. Discussion on gathering skin conductance data

When recording skin conductance, there are some factors that may affect the signal, as can be external temperature and humidity as well as the movement of the sensor that may induce some noise on the signal [115]. In addition, this signal can also be influenced by the intake of some medications that can change the sweating levels.

4.2.1.c.iii. Skin temperature

The information that can be collected from the skin temperature and the issues to be considered in the process is described next.

4.2.1.c.iii.1. Data description

Skin temperature is a signal quite similar to the skin conductance, highly influenced by the sympathetic nervous system, so it also reacts to fight-or-flight situations.

Just like GSR, as it is a continuous signal with no patterns to process, the processing to be performed is the extraction of indicators such as the mean value, range, etc.

4.2.1.c.iii.2. Discussion on gathering skin temperature data

The signal may reflect changes due to external factors when measuring it such as external temperature, humidity or movement that may generate noise in the signal. In addition, it can also be influenced by consumption of food, alcohol, weight-loss diet, physical activity, etc.

4.2.1.c.iv. Breathing rate

The information that can be collected from the breathing rate and the issues to be considered in the process is described next.

4.2.1.c.iv.1. Data description

Respiratory system has relation with the parasympathetic nervous system, having a strong influence on relaxation [45]. Depending on the devices available, different features can be measured. Some studies measure the concentration of some substances from breathing, such as FetCO2 [61]. Other common devices consist of a belt that measures the lung volume change as it inhales or exhales air.

From this device, many indicators can be used, such as breathing capacity used, number of inspiration/expirations or number of nouns, but this would need a highly detailed noise reduction work before extracting these two last features (to remove talking noise, sneezes, coughs, etc.).

4.2.1.c.iv.2. Discussion on gathering breathing data

This signal may be influenced by pressing the sensor against the back of the chair, by speaking, coughing, sneezing, etc.

4.2.1.d. ***Facial expressions***

As commented in the review of the state of the art, another common adopted approach to detect emotions is the detection of facial gestures. However, within the aDeNu group, this research is led by one of the psychologists and in addition, the analysis of this information is part of the work carried out by our colleagues of the MAMIPEC project from the University of Valencia. Thus, facial gestures detection is not part of the research of this Thesis. Nevertheless, for completeness of the approach proposed in this work and the description of the experiment carried out as well as to understand some of the design decisions, there will be some mentions along the text.

The link between facial gestures and emotions has been studied for centuries (even Charles Darwin studied this approach in The Expression of the Emotions in Man and Animals), but the most popular approach during the last three decades has been the developed by Paul Ekman, who found that there are 6 universal expressions with a high agreement regardless the cultural background: anger, disgust, fear, happiness, sadness, and surprise [73]. Other alternatives to detect facial gestures can be used such as electromyography, which can offer really detailed values on measuring certain facial muscles activity. However, it is a more intrusive technique as it needs to attach detectors to the skin.

The precision of the measurements may vary depending on several factors, some of them technical related to the device used and other not technical (lighting, device position, movement of the learner’s face) which may difficult the task of processing the video files recorded to get indicators from the expressions registered.

When capturing facial expressions, some information relative to the subject should be known that may determine certain aspects of the facial behavior to be registered with that user. One case, for instance is, when gathering facial expressions from blind people, some “brusque” head movements may be done in case they are using speakers, looking for the best angle to receive the audio stream, these movements are called blindisms.

In the experiment described in this work, facial expressions were recorded both by using a webcam and a Kinect for windows device.

4.2.1.d.i. Discussion on gathering webcam data

In addition to the aforementioned issues when capturing facial gestures detection, the quality of the video file depends of several factors as the sensor of the camera, the configuration of the recording quality and the processing of the video when saving it (depending on the codec used, the compression level, etc.). Usually, a good quality video takes a lot of space in hard disk and a lot of computational resources, so the space available and the computational resources should be taken into account when deciding the technical specifications of the device to use and the compression to apply.

4.2.1.e. *Sentiment analysis*

Another growing data source used for emotion detection is the sentiment analysis. It is considered in our multimodal approach to be performed with the outcomes of emotional reports that can be collected from the participants.

The emotional reports consists of collecting information about relationships between the emotions the participants feels and its impact on the learning strategies. In particular, the subject expresses the way he or she felt while solving a given task. The use of this resource relies on the basis that emotions can have an effect on the cognitive process by initiating, accelerating, altering or interrupting it [170]. The impact of emotions on the cognitive process influences learning strategies [118]. Due this fact, the sense of the emotion (e.g. positive or negative) felt by the user along the problem solving process can determine the learning strategies selection, application and effectiveness in order to solve it. In this sense emotions impact on the user behavioral, motor and physiological responses, learning from mistakes, decision making, storing and retrieving relevant information [91].

4.2.1.e.i. Data description and gathering

Nowadays, there are several approaches for gathering the emotional information from the learner. We used free text forms provided through the learning management system interface where the learner was asked (after finishing a task) to fulfill the following four statements: 1) “While performing the task I felt…”, ii) “While performing the task I thought…”, 3) “The difficulties encountered in order to solve the task have been…”, and iv) “I solved these difficulties by …. “. These questions were defined by a psycho-educational expert.

This information can be processed in several ways [30] depending on the abstraction level, from analyzing texts from a bag of words approach, evaluating each one regardless the possible relations it could have with other words, to more complex techniques trying to extract the affective charge of sentences.

As a first approach, and since we do not have previous experience on sentiment analysis, an affective database can be used to carry out a sentiment analysis on the text and counted the terms with positive valence and negative valence, producing a similar categorization as the expert (positive, negative, neutral and ambivalence).

The result of this processing would give us an emotional score for each text typed by the participant based on the ratio of positive and negative terms used in the text.

The emotional reports were not only collected to perform sentiment analysis, but also to provide user-defined labels to the data collected from the other data sources which allow to apply the supervised data mining techniques to that data, as discussed below.

4.2.2. Labeling

As reported in the corresponding section in the state of the art review, the data labeling is one of the aspects were more diversity can be found in affective computing works. From the identified factors, a proposal to be followed on the experiment is here presented.

For the current work, the approach to be followed consists of a dimensional approach using the valence and arousal dimensions (the dominance was discarded due to its complexity to be understood and evaluated). This way, the emotions labeled can be grouped in different categories when predicting it (being able to adapt a categorical approach.

The experiments are to be designed for the participant to be the one to label his or her emotions while the experiment is being carried out by means of the SAM scale [39] after each one of the tasks to be proposed to solve during the experiment. This approach is the most followed approach seen in the works viewed in section 2, but some experts are asked to also label the data after the experiment. An open labeling approach during the experiment has also been chosen to be used as an alternative way of getting information, asking the participant to type the emotions felt after solving several tasks. Thanks to that open labeling approach (the emotional reports) we can have another data source from processing those texts using the sentiment analysis.

4.2.3. Task

Choosing the task to be performed is an important point as our goal is to elicit affective states in the educational domain. With that in mind, the Mathematical subject was chosen as appropriate due to several issues. First, a large number of students have been found to have negative feelings about Mathematics [133]. Negative emotions toward Mathematical tasks have been explored in different cultures and can be detected in different groups of age, gender, educational systems, etc. Many people have developed negative attitude towards Mathematics showing negative emotions during problem solving situations which are interfering in their cognitive process in a negative way [58]. People who experience negative emotions toward Mathematical tasks can suffer from, all, or a combination of the following situations: difficulty in thinking, feelings of panic, tension, helplessness, fear, shame, nervousness and loss of ability to concentrate, negative self-talk, and/or a general sense of uneasiness [15,172,226,235]. These negative emotions are distressing in itself and also tend to impair Mathematical performance [16] because emotion and cognition are seen as two complementary aspects of mind. In this sense, Mathematical subject is an optimal educational issue to cause emotions.

4.2.4. Emotion Elicitation Method

Once the task has been already chosen, the way to elicit emotions has to be designed. As aforementioned, the Mathematics field is commonly associated with negative affective states. In order to elicit those affective states, a series of strategies, associated with the task, are to be used. The variables to tweak in order to elicit those affective states are: i) difficulty (by introducing severe changes in the difficulty of the task, as these changes may confuse and frustrate the participant in case the difficulty is abruptly risen or pleasing the participant in case it is lowered or even bore the participant in case the difficulty is set to an excessively low level) and ii) time limits (by introducing countdowns in some tasks in order to introduce stress to the participant when dealing with the proposed task).

4.2.5. Labeling Approach

One of the main methodological variables in our experiment is the approach to be followed in the labeling process. As we aim to use supervised learning techniques, a good labeling is crucial in order to get a good system. This is because the importance of the training (labeled) data instances in the model generation. In this sense, we are going to compare different sources of labeling in order to evaluate how those differences may impact on the results of the model generated. In this case, different external annotators will provide different emotional labels and models will be generated with each one of the labeling approaches followed in order to see which approach provides better results.

4.2.6. Model Generation algorithm

Another open point found in literature is the technique to be used in the model generation process. As seen in Table 2, most works use several algorithms in order to evaluate which one provides the best results. In the work here proposed, that approach is also to be followed, using some of the most common algorithms used in literature and comparing their results.

4.3. Context

During Madrid's Science Week in 2012[5], four activities were proposed and carried out by the aDeNu research group as part of the research works of the MAMIPEC project with two main purposes. On the one hand, to create a dataset of affective information collected from multiple data sources when certain Mathematical tasks are carried out in order to train different prediction systems. On the other hand, following the Madrid Science Week goal, allow the Madrid citizens to know what we as researchers do in our laboratory and show them where affective computing and e-learning are going.

The activities were announced in the pamphlet distributed all over Madrid with the activities to be hold during the Madrid Science Week (which really consists in a two-week period) giving enough information to know what was the activity about but not too much in order to avoid the people coming to be prepared for an emotional experiment. The four activities designed were the following:

[5] https://adenu.ia.uned.es/web/es/Proyectos/Semana%20Ciencia/2012

- *Activity 1: What do you feel when solving mind games? Would you dare to create one?* This activity was different to the other three proposed. This one relies on the use of a collaborative platform (called the Collaborative Logic Framework) to solve a mind game and propose a new one in groups up to 4 people.
- *Activity 2: Logic reasoning capability: Which role do emotions play on it?* This activity consisted on a series of 3 blocks of Math problems for university students.
- *Activity 3: Ambient Intelligence: Affective automated tutor for the "everyday mathematics".* This activity had the same structure than the previous one but was designed for the general public.
- *Activity 4: Improving abstraction skills through problem solving and teamwork.* This activity was like activities 2 and 3, but oriented to high-school students.

The experimental conditions differed between activity 1 and the other three. In particular, the first one was a collaborative activity and the other three were individual ones. Research on collaboration and affective issues related to the activity 1 was carried out in another Master Thesis [139]. In the current work, the research focuses on how to detect emotions in individuals learning by their own. Therefore, activities 2, 3 and 4 where participants are solving Mathematical problems individually, were designed with this goal in mind, as commented below. The purpose of having three different activities addressed to different profiles allows gathering a more heterogeneous sample of participants.

In order to get as many participants as possible, several sessions (scheduled in periods of 2 hours) were carried out. When participants registered for the activities in an online form, they indicated the sessions for which they had availability. Up to four participants could carry out the individual activity in each session as four individual stands were configured in the aDeNu laboratory. Each stand was separated from the rest with panels, so participants could not see each other's computers and could carry out the activity on their own.

4.4.Participants

The participants of the experiment were 78 people (43 males and 35 females) with an average age of 25.5 years (with a standard deviation of 12.4). The average height was of 169.4 (standard deviation 8.91) and an average weight of 62.8 (standard deviation 12). 42 participants said that they practiced sport regularly and 28 admitted to suffer stress situations in the previous days of the experiment.

From the psychological questionnaires (BFI, GSE and PANAS), results (average and standard deviation for the corresponding indicators) are shown in Table 5 (BFI), Table 6 (GSE) and Table 7 (PANAS).

	Extraversion	Agreeableness	Conscientiousness*	Neuroticism*	Openness to experience*
Average	32	36	30	22	37
Standard deviation	9.9	7	6	6	7

Table 5. Average and standard deviation of the BFI questionnaire.

Those features marked with an asterisk in Table 5 were not calculated for underage participants.

	General Self-Efficacy
Average	36
Standard deviation	9

Table 6. Average and standard deviation of the GSE questionnaire.

.

	Positive aspect	Negative aspect	Affect Balance Scale
Average	33	17	16
Standard deviation	7	7	8.4

Table 7. Average and standard deviation of the PANAS questionnaire.

4.5. Design

The design of the experiment was carried out with the support of three psychologists[6] with a strong background on psycho-educational and psycho-emotional issues. Nevertheless, the technological decisions and deployment was led by this Ph.D. Thesis. Next, the infrastructure, materials, implementation and structure are described.

4.5.1. Data sources

The data sources used in this experiment where chosen from those evaluated in the review of the literature performed in section 2.1. Here are the details on how the different data sources were set up in the experiment:

4.5.1.a. *Keyboard*

A keylogger/mouse tracker application was developed as part of this experiment to extract all the interactions with the keyboard and mouse. Every key event registered by the application has to be stored with a timestamp as precisely as possible, so the system time (including milliseconds) is stored together with the event information (i.e., if it has been a press or a release event and the ASCII code of the key pressed).

[6] Mar Saneiro, Pilar Quirós and Raúl Cabestrero

4.5.1.b. *Mouse*

To log all the interactions with the mouse, the aforementioned keylogger/mouse tracker application developed was also used, as the software can get the mouse interactions (except for the scroll interactions).

4.5.1.c. *Physiological signals*

The bio-feedback device J&J Engineering I-330-C2 system[7] was used to record the following physiological signals: Heart rate, breath volume, skin conductance and skin temperature. This device, powered by 4 AAA batteries, has a USB connector so it can transfer the data to a computer. It has also two input ports where different measuring devices (such as electrodes for ECG, chest belt for breathing, etc.) can be connected.

The device is distributed with a recording software (that requires Windows from version 98 to XP) called Physiolab. The software is proprietary and does not allow exporting the data recorded live. It only allows exporting the data once the recording has finished, and the data can be exported as an Excel file or an ASCII file, being this last one a csv-like (comma separated values) format, separated by tabulators. It is important to highlight two things when exporting the data recorded by this software: 1) the timestamps are relative to the moment the recording has begun (starting all the recordings with the value 0:00.000) instead of recording the system time (which would allow to synchronize the signals recorded with other devices, as discussed later), and 2) instead of having one single column indicating the time the row values were recorded, there is one time column per each signal column (having many duplicated columns), and there are values generated with different frequencies. On the one hand, heart rate, breathing, skin temperature and skin conductance signals are recorded every 100 ms. On the other hand some values computed from the Discrete Frequency Transform over the ECG data are generated every 500 ms (as they are computed from other signals recorded with a higher frequency). This means there are many rows that have values registered in different times, which causes inconsistencies in the data of a row as it contains values from two different moments.

The I-330-C2 allows collecting the afore-described signals as follows. The HR is measured by placing three electrodes on the participants (one in each ankle and another in the chest, over the heart) our device records the heart rate every 100 ms. The GSR is recorded by two sensors attached to two velcro straps to be placed on the index and ring fingers of the non-dominant hand in order to avoid the movement of the hand when moving (if using) the mouse. The skin temperature is recorded by placing a sounding fixed to the wrist using a bracelet. Finally, the breathing rate is recorded by placing a belt around the learner chest which measured the respiratory volume oscillations. All the signals were recorded every 100ms.

4.5.1.d. *Facial expressions*

In order to record the facial expressions, the most common device used to this end was used: a webcam. Due to the affordability and quality of webcams nowadays,

[7] http://www.jjengineering.com/C6.htm

webcams can be used to record the participants face. This device usually produces a file containing the video, which should be processed via artificial vision techniques.

Nevertheless, during last years, new devices have appeared with added sensors to provide extended image recording capabilities. This is the case of the Kinect. Kinect is a device originally released in 2010 as a console controller (for the Microsoft’s Xbox 360 console), based on a camera and a depth sensor (although it also includes an array of microphones), able to record not only a video of the objects in front of it but also the depth they are at (using a matrix of infrared beams). This way, they can reproduce a pseudo 3D recreation of the recorded scene, limited to one point of view (not creating a 3D model of the recorded objects, creating a 3D model of the side of the objects recorded).

This device was chosen to be used as Microsoft (the company behind Kinect) released a computer compatible version of Kinect with face tracking capabilities. This feature allows to detect[8] the position of 100 characteristic face points as weights of six Action Units (Neutral, Upper Lip Raiser, Jaw Lowerer, Lip Stretcher, Brow Lowerer, Lip Corner Depressor, Outer Brow Raiser) and 11 Shape Units (Head height, Eyebrows vertical position, Eyes vertical position, Eyes width, Eyes height, Eye separation distance, Nose vertical position, Mouth vertical position, Mouth width, Eyes vertical difference, Chin width), which are a subset of what is defined in the Candide3 model.

4.5.1.e. *Sentiment analysis*

In order to perform sentiment analysis, no additional devices are required. In contrast, in this experiment, participants were asked to type how they felt during the different tasks proposed. The idea is to analyze the text from the participants in order to extract affective information in an automated way.

4.5.2. Labeling

As one of the methodological variables to study in this experiment is the information to use to label the affective state of the participants during their interaction. In this sense several ways of labeling were designed, depending on the labeler and the time those labels are to be generated. In order to allow the participants to express their affective state, two different emotional ways of reporting their emotions were included in the experiments: i) the Self-Assessment Manikin [39] (shown in Figure 6), was included after every single problem allowing the participants to indicate the valence and arousal dimensions of their affective state in a 9-point Likert scale for each one of the dimensions included and ii) a text area was shown after every set of problems asking the participants to express their emotions. From the data collected, different affective labels are to be generated so the machine learning algorithms can generate models to perform predictions according to those labels.

[8] http://msdn.microsoft.com/en-us/library/jj130970.aspx

4.5.3. Tasks

As mentioned in section 4.2.3, Mathematics was chosen as the subject to propose tasks to our participants. To select the materials to be used, a series of mathematical problems was chosen from repository of mathematical problems provided by the BBC. Two psychologists[9] selected and classified problems according to their difficulty. Graphical logical series were also selected to create a final task.

4.5.4. Infrastructure

For the experiment four stands were set up, so we could host four participants per experiment session. Each stand consisted of four computers and a tutor supervising the activity in each stand. Table 8 shows the configuration of each stand.

First, we had the computer were the participant carried out the tasks through a web browser. In addition, there was some software to i) record the screen (to facilitate the analysis of the interactions after the experiment), ii) show the screen in another computer so the tutor could see what the participant was doing without disturbing him or her, and iii) to collect data with affective information (i.e., the keylogger/mouse tracker application developed). Another computer was used to run the bio-feedback equipment and record the physiological signals. It also recorded a video of the participants' face. A third computer was used by the tutor to remotely see the participant's screen. Finally, there was a fourth computer that recorded information from the Kinect device.

Computer	Used by	Running software	Devices attached
Participant's computer	Participant	• Web browser (Google Chrome): in order to use the e-learning platform • Screen recorder (CamStudio Portable): to record the participant's computer screen. It should be hidden in order not to disturb the participant or make him or her feel monitored. • Keylogger/mouse tracker app (implemented): to record the participant's interactions with keyboard and mouse. • Remote desktop program (VNC): to allow the participant's tutor be aware of the advances of the participant during the experiment, allowing him or her to take a timestamp every time the participant ends a task.	none

[9] Pilar Quirós and Raúl Cabestrero

Computer	Used by	Running software	Devices attached
Webcam / Physiological recording computer	Tutor	• Webcam recorder (Logitech webcam software): to record the participant's face • Physiological signal recorder (Physiolab): to record the participant's physiological signals and take the timestamps to know the time every task has started and ended.	• Webcam (Logitech C310 or Quickcam Pro 9000) • Physiological sensors (J&J Engineering I-330-C2)
Desktop monitoring computer	Tutor	• Remote desktop program (VNC): to allow the participant's tutor be aware of the advances of the participant during the experiment, allowing him or her to take a timestamp every time the participant ends a task.	none
Kinect recording computer	Tutor	• Kinect video recording program (Kinect studio): in order to save the data recorded by the Kinect device. • Kinect facial data exporter (developed by Miguel Arevalillo from Universidad de Valencia): to export the captured data points into a csv file live during the experiment.	• Kinect device • External hard disk

Table 8. Configuration of the computers used in each stand in stage 1

In order to synchronize all the signals, an additional computer was set up as time server for all the stands, so all the computers used in the experiment in each stand synchronize their time to the time signal provided by the time server. To make this possible, all the computers were connected under the same network. Synchronization of the information collected is critical when adopting a multimodal approach in order to guarantee that the possible detected reactions on the user from the different data sources correspond to the same event in a concrete time point.

As webcam and Kinect devices are used to record the participant's face, when placing the computers, a layout has to be designed to allow the devices capture the participant's face. To allow that, the computers were placed as shown in Figure 13 and Figure 14.

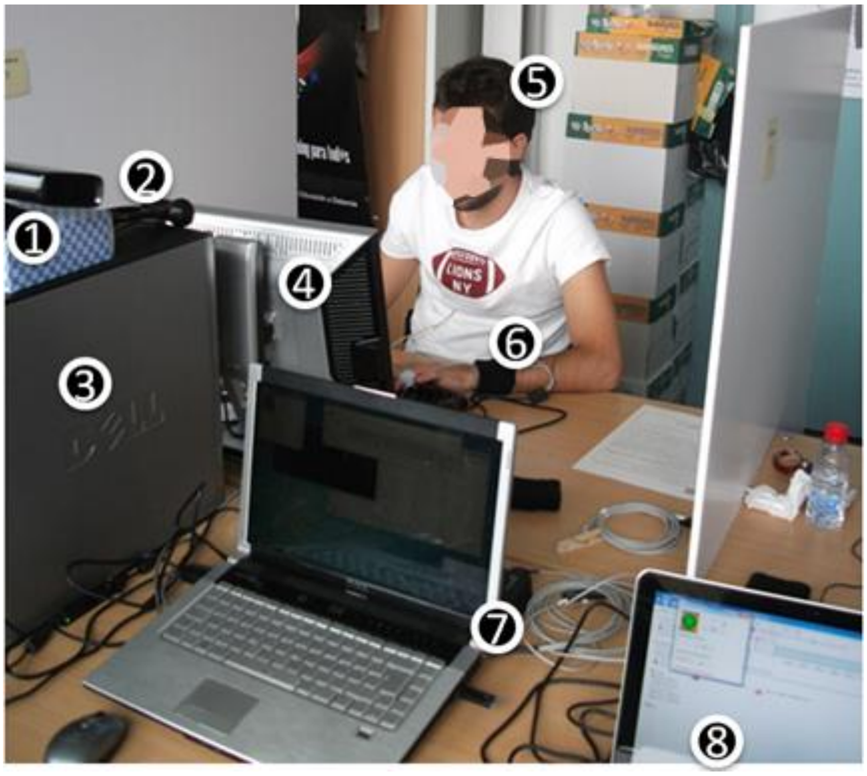


Figure 13. Stand configuration from the tutor perspective

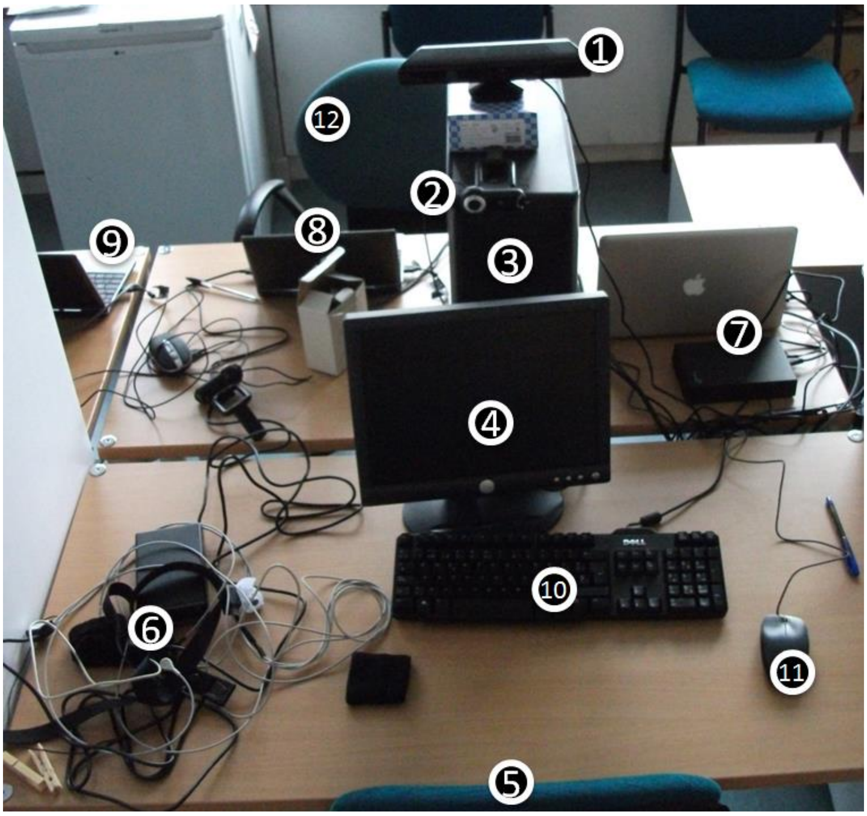


Figure 14. Stand configuration from the participant perspective

Figure 13and Figure 14 offer, respectively, the point of view of the tutor and the participant, where the following components are shown: 1) Kinect device, 2) webcam,

3) participant's computer, 4) participant's screen, 5) participant's position, 6) Physiological devices, 7) Kinect recording computer, 8) webcam/physiological recording computer, 9) desktop monitoring computer, 10) participant's keyboard, 11) participant's mouse and 12) tutor's position.

Despite the technical resources, as aforementioned, a tutor was needed per stand in order to support the participant during the experiment and follow the deployment of the technological infrastructure issues during the experiment. In particular, the tutor i) guided the participant through the session, helping and guiding him or her in case it is needed, ii) provided the participant the questionnaires to fill up (as described in the Materials' section), iii) attached the bio-feedback sensors to the participant, iv) took timestamps to get the physiological data labeled, and v) detected if something went wrong and if possible, corrected it.

During the experiments, there were also two more persons in the laboratory, as far away as possible from the stands in order to avoid distracting participants. These two people were: i) a Master of ceremony, responsible for welcoming and talking to all the participants, who gave the initial instructions for the experiment and orchestrated the session, and ii) a technician expert, to prepare the technological infrastructure for each session, take actions if possible in case some device fails during the experiment, and save after the session the data recorded. When possible, a psycho-educational expert was watching the participants and taking notes of their body movements. When this was not possible, the participants were recorded with a video camera so the movements.

4.5.5. Materials

The materials prepared to perform this experiment are listed next.

- ***Information consent:***
  As the participant's face are recorded during the experiment, he or she had to be informed of the use of that information within the MAMIPEC project and sign their agreement to allow us recorded and use the data in those terms. Within the project, we also guarantee that the data is stored in a secure way in order to avoid possible data leaks. When participants were visually impaired, accessible electronic versions were provided.
- ***Demographic and psychological questionnaires:*** Some questionnaires to get some demographic information, and psychological information from the participants and their personality were asked to be filled by the participant.
- ***Calibration questions:***
  A series of questions selected to evaluate the participant's physiological reactivity. These questions included simple questions (e.g. is Paris the capital of France?) and awkward questions (e.g. have you ever taken something from a store without paying it?) in order to see if the awkward questions made the physiological signals change.

- ***Calibration images:***

  Eight images, extracted from the International Affective Picture System (IAPS), to elicit emotions and see their impact on the physiological signals. The last two images were explicitly strong to trigger a sudden change in the participant. For the activity 4 (oriented to high-school students), the strongest images from our choice were replaced by other less strong images, as participants could be under 18 years old, and thus, not appropriate for them.
- ***Calibration sounds:***

  In order to check the physiological changes in visually impaired participants, sounds were used instead of images. These sounds were picked from the International Affective Digitized Sounds (IADS) and included sounds from a yawn to a feminine orgasm, as the purpose was similar to the calibration images and changes in the participants' physiological reactions were sought.
- ***Math exercises:***

  A series of Math exercises were categorized depending on their difficulty. The problems were chosen from a repository provided by the BBC taking into account that they were going to be solved with no paper to perform the mathematical operations, so they had to be solved mentally.
- ***Graphical logical series:***

  Logical series were chosen for the third and last activity, looking for offering a low difficulty level so the participant ended the experience with a comforting feeling. For those visually impaired participants, an alternative was chosen, based on alphanumerical logical series.
- Satisfaction questionnaire and PANAS:

  For finishing the experiment, the PANAS questionnaire was elected to be fulfilled, as a free text to tell us about their opinions about the experiment.

The contents of all the materials here presented are included in Appendix II (section 13.2).

4.5.6. Implementation

The technological infrastructure prepared for the experiment included the keylogger and mouse tracker application and the configuration of the Mathematical tasks in a learning management system.

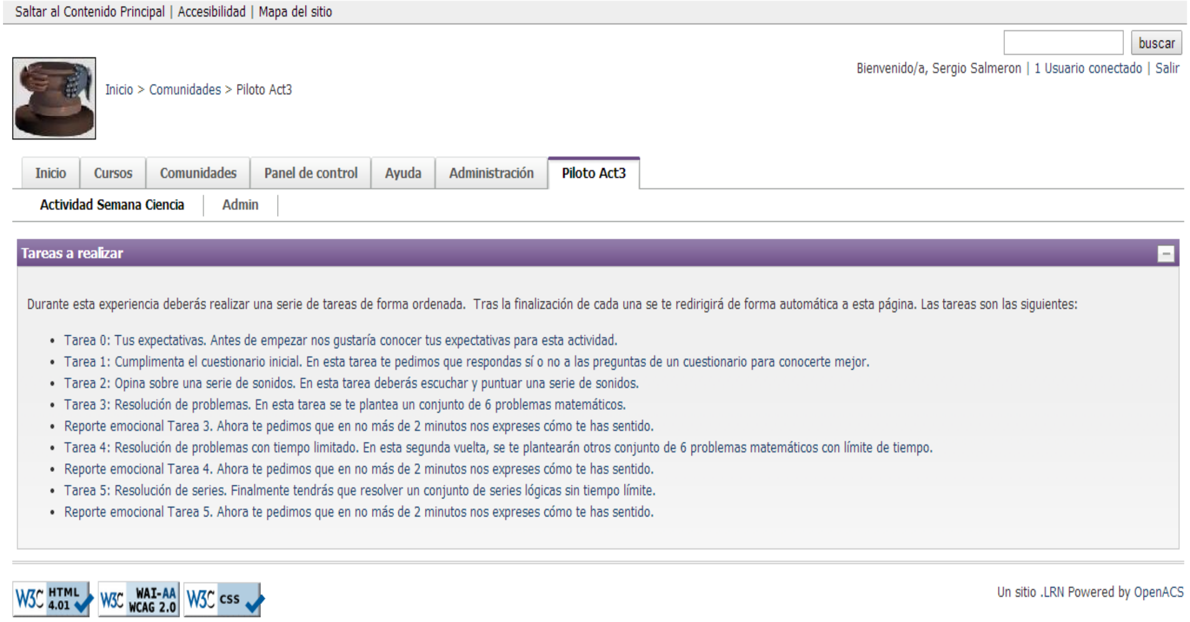
4.5.6.a. *Keylogger and mouse tracker application*

As commented in the Section 4.5.1.a and 4.5.1.b, a keylogger/mouse tracker app was developed in Java to collect all the interactions with keyboard and mouse. To do that, all the events triggered by these devices need to be collected. As all the events (and the app was going to be invisible in order not to disturb participants) need to be collected, the application developed had to communicate with the operative system in order to get all these interactions, so an external library was used. The library found capable to offer information from all the interactions was the Java System Hook by Ksquared.de, which offers an easy way to access, via Java, to all the information required.

The application was developed with no graphical user interface. When ran, it automatically creates two log files (one for the mouse interactions and one for the keyboard interactions) and starts to write all the events recorded, each one with its corresponding timestamp.. In addition, it was added the functionality to export the active process during each event. Some known issues currently open are: the app is only compatible with MS Windows devices and in case several monitors are being used, it does not detect in which monitor the event happened.

4.5.6.b. *Learning management system*

Regarding the learning management system, dotLRN was used as it is well-known for its adaptive and accessibility capabilities [210]. Thus, the environment used during the experiment for the participant to interact with was developed in a dotLRN server. All of the tasks were implemented as dotLRN assessments, showed ordered in an initial splash screen (see Figure 15) where participants were redirected when finishing every task. Figure 16 shows one of the Mathematical tasks as displayed for the experiment.

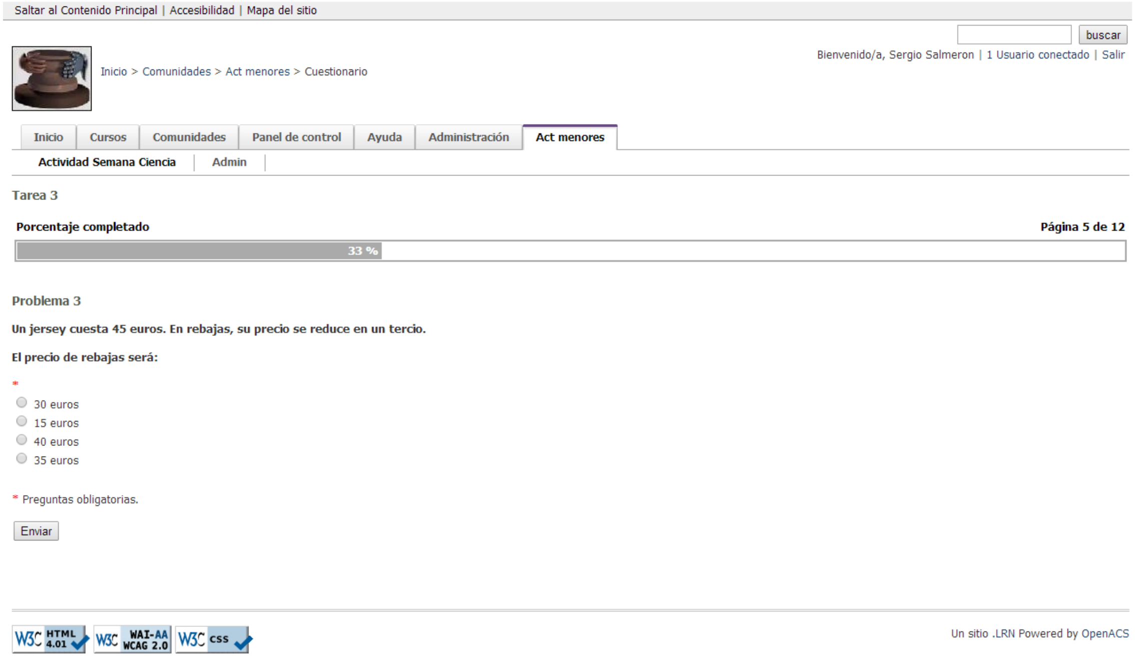

Saltar al Contenido Principal | Accesibilidad | Mapa del sitio
buscar
Bienvenido/a, Sergio Salmeron | 1 Usuario conectado | Salir
Inicio > Comunidades > Piloto Act3
Inicio
Cursos
Comunidades
Panel de control
Ayuda
Administración
Piloto Act3
Actividad Semana Ciencia
Admin

Tareas a realizar

Durante esta experiencia deberás realizar una serie de tareas de forma ordenada. Tras la finalización de cada una se te redirigirá de forma automática a esta página. Las tareas son las siguientes:

- Tarea 0: Tus expectativas. Antes de empezar nos gustaría conocer tus expectativas para esta actividad.
- Tarea 1: Cumplimenta el cuestionario inicial. En esta tarea te pedimos que respondas sí o no a las preguntas de un cuestionario para conocerte mejor.
- Tarea 2: Opina sobre una serie de sonidos. En esta tarea deberás escuchar y puntuar una serie de sonidos.
- Tarea 3: Resolución de problemas. En esta tarea se te plantea un conjunto de 6 problemas matemáticos.
- Reporte emocional Tarea 3. Ahora te pedimos que en no más de 2 minutos nos expreses cómo te has sentido.
- Tarea 4: Resolución de problemas con tiempo limitado. En esta segunda vuelta, se te plantearán otros conjunto de 6 problemas matemáticos con límite de tiempo.
- Reporte emocional Tarea 4. Ahora te pedimos que en no más de 2 minutos nos expreses cómo te has sentido.
- Tarea 5: Resolución de series. Finalmente tendrás que resolver un conjunto de series lógicas sin tiempo límite.
- Reporte emocional Tarea 5. Ahora te pedimos que en no más de 2 minutos nos expreses cómo te has sentido.

Figure 15. List of tasks to be done by the participant


Saltar al Contenido Principal | Accesibilidad | Mapa del sitio
buscar
Bienvenido/a, Sergio Salmeron | 1 Usuario conectado | Salir
Inicio > Comunidades > Act menores > Cuestionario
Inicio
Cursos
Comunidades
Panel de control
Ayuda
Administración
Act menores
Actividad Semana Ciencia
Admin
Enviar
Un sitio .LRN Powered by OpenACS

Tarea 3

Porcentaje completado — Página 5 de 12

33 %

Problema 3

Un jersey cuesta 45 euros. En rebajas, su precio se reduce en un tercio.

El precio de rebajas será:

*

○ 30 euros
○ 15 euros
○ 40 euros
○ 35 euros

* Preguntas obligatorias.

Figure 16. Sample of a Mathematical problem proposed.

The implementation of the problems in the platform was carried out carefully, taking into account accessibility all time, taking advantage of the accessibility of the platform itself. Even the calibration images task was subtitled when possible, or substituted for other based on affective audios when appropriate.

4.5.7. Procedure

The experiment was structured in three different parts, with different tasks as shown in each of the boxes in Figure 17.

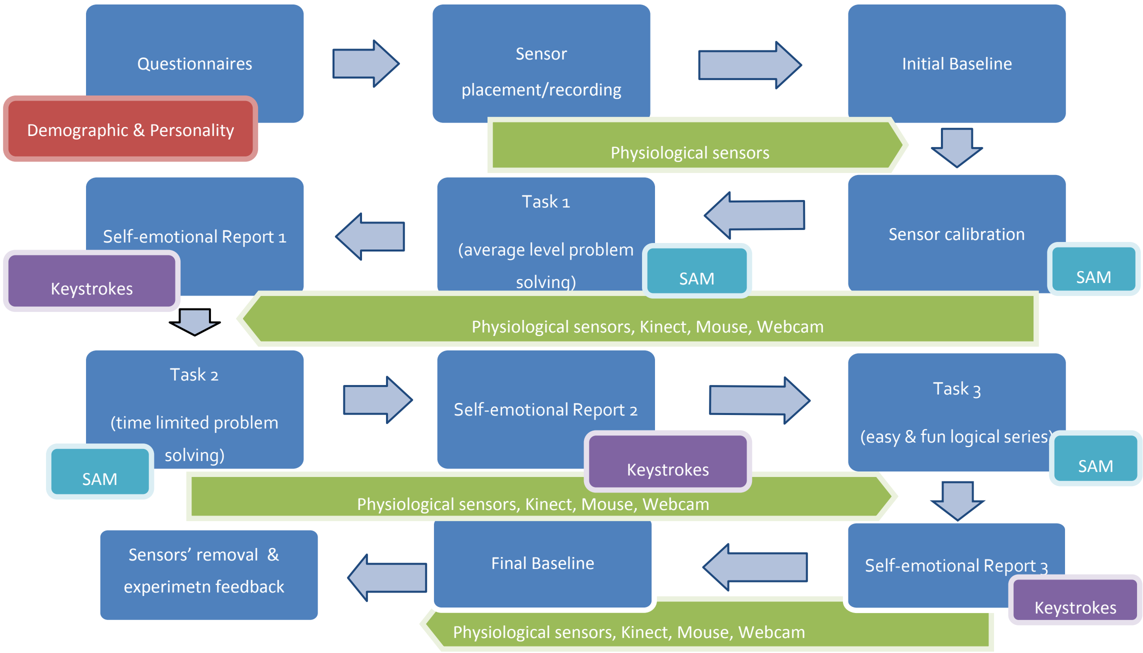

Questionnaires
Demographic & Personality
Sensor placement/recording
Initial Baseline
Physiological sensors
Sensor calibration
SAM
Task 1 (average level problem solving)
SAM
Self-emotional Report 1
Keystrokes
Physiological sensors, Kinect, Mouse, Webcam
Task 2 (time limited problem solving)
SAM
Self-emotional Report 2
Keystrokes
Task 3 (easy & fun logical series)
SAM
Physiological sensors, Kinect, Mouse, Webcam
Self-emotional Report 3
Keystrokes
Final Baseline
Sensors' removal & experimetn feedback
Physiological sensors, Kinect, Mouse, Webcam

Figure 17. Stage 1 experimental structure, including tasks and data to be collected.

4.5.7.a. *Part 1: setting up*

The first part of the experiment consisted of a series of tasks designed to set up and calibrate the recording devices to be used. Four blocks are considered:

Block 1. Questionnaires fulfillment. Participants had to fulfill a series of questionnaires before the experiment began:

- Demographic information: a general questionnaire in order to collect information about gender, age, computer skills, information that may affect some sensors measures (sports, smoking, medicines) and possible allergies (to avoid using the latex electrodes with someone allergic to latex).
- Big Five Inventory (BFI): a 44 item questionnaire to extract 5 dimensions of personality: extraversion (sociability), neuroticism (tendency to experience negative or unpleasant emotions easily), conscientiousness (tendency to be organized), agreeableness (tendency to be friendly) and openness to experience (curiosity and lack of uncomfortableness for new things) [24].
- General Self-Efficacy (GSE): this 10 item questionnaire provides information about the expectations of the ability to face any difficult situation [222,223].

Block 2. Sensor placement and recording. All the recording devices were set (if not running yet) to start recording:

- Heart rate sensor: Participants were attached three latex electrodes in order to record their heart rate. Two electrodes were set on the inner side of the ankles and the other one on the chest over the heart.
- Respiratory sensor: A belt was tied around the participant’s chest in order to registry the volume of air consumed.
- Skin conductance sensor: Two velcro straps to be placed on the index and ring fingers of the non-dominant hand of the participant.
- Temperature sensor: A sounding was placed in contact to the participant’s wrist attached by using a wristband.
- Screen recording: The program CamStudio was configured and started recording before the participant entered the room.
- Mouse tracker: The program developed to record mouse interactions was launched before the participant entered the room.
- Key logger: The program developed to record keyboard interactions was launched before the participant entered the room.
- Remote desktop: The program VNC server was set up to allow the participant’s tutor view the participant’s screen all along the experiment.
- Webcam: The webcam software used (provided by Logitech with the webcams) started recording after the physiological sensors were placed
- Kinect: Kinect started recording after the physiological sensors were placed. A program developed by Universidad de Valencia to export live the facial points detected by the applications was also launched at this time.

Block 3. Initial base line. Participants were asked to relax for 2 minutes in order to get the values of their physiological signals while relaxed.

Block 4. Sensor calibration. Some calibration questions and images or sounds were used.

- Calibration questions: A set of 7 questions were asked to see the signal changes when they were asked awkward questions.
- Calibration images / sounds: 8 Images extracted from a standardized affective image database were shown in order to see the participants' reactions to them. In case the participant is visually impaired, during this task, 8 sounds extracted from a standardized affective sound database were played. At the beginning of this task, participants were explained the Self-Assessment Manikin (SAM) scale so they could score the images/sounds using that scale.

4.5.7.b. ***Part 2: task solving***

The second part corresponds to the activities proposed during the Mathematical task. Here starts the experiment itself, where the participants are dealing with an e-learning platform to perform the mathematical tasks. Three group of tasks are carried out, with the same structure. First, a task with 6 problems (from those described in section4.5.5) is done. Next, the emotional report explained in section 4.2.1.e is asked to fill in. In them, participants were asked to type their feelings while solving the problems. They had no time or space limit to express themselves.

Regarding the tasks, they were design as follows:

- Task 1. Problem solving. A set of 6 problems with a low-medium difficulty level had to be solved.
- Task 2. Problem solving with time limit and higher difficulty. Before starting this task, participants were informed that there was a 3 minute time limit in that task. The 6 problems in this case were more difficult than the ones presented in the previous task (but participants were told that these problems used to be solved much faster than the previous ones in order to generate a contrast between the low difficulty level expected in this task with the real high difficulty found). In this task we expect to elicit stress and frustration in our participants.
- Task 3. Graphic logical series. A series of 6 easy graphic logical series where given to participants to be solved. During this task we expected participants to feel better than in the previous task due to the low difficulty of this task.

4.5.7.c. ***Part 3: experiment ending***

The third and last part ended the experiment, and collected the participants' baseline at the end (i.e., participants are asked again to relax for 2 minutes), removed the sensors from them and asked them some feedback with the following questionnaires:

- Positive and Negative Affect Schedule (PANAS): a 10 item questionnaire to measure the primary dimensions of the mood [203,243].

- Satisfaction questionnaire: to allow participants report if they liked the experiment.

Once the experience ended, participants were debriefed about the experiment and allowed to ask as many questions as they wanted. They were also shown how the information was recorded and why.

4.6. **Data recorded**

Once the experiment ended, it was time to prepare the data to be processed. To match all the data from the different data sources, the following labeling was used to save each participant log/recording files:

act*M***usr***N***ses***OO***d***PP***m***QQ*

Being:

- M: the number of the activity ({1,2,3 or 4})
- N: the number of the stand the participant was seated on ({1,2,3 or 4})
- OO: the session (starting time) of the experiment that participant took part in ({00-23})
- PP: the day the participant came to participate in the experiment ({01-31})
- QQ: the month the participant came to participate in the experiment ({01-12})

The data files generated for each participant included information from the different devices used in the experiment. Details are provided next.

4.6.1. Webcam video

The result of the webcam recording was a .wmv file containing the video with a 1280x720 resolution and 15 frames per second. As two different webcam model were used, the files generated were different.

- Files recorded with Logitech C310 had a 2 channels 48 kHz audio track, generating a data stream of about 30-50 mb/min
- Files recorded with Logitech Quickcam Pro 9000 had a 1 channels 32 kHz audio track, generating a data stream of about 3-6 mb/min

As aforementioned, this information was not used in the research work reported in this Thesis, but is included here for completeness and to support future work.

4.6.2. Kinect video

When using the Kinect Studio during our experiments, a video file was generated, containing the image and the depth data. The output is a .xed file with a 640x480 resolution at 20 frames per second, requiring each file near 1 gb/min. This file can be opened only by Kinect Studio when a Kinect device is connected.

As aforementioned, this information was not used in the research work reported in this Thesis, but is included here for completeness and to support future work.

4.6.3. Kinect facial points

A csv file containing all the facial information provided by the Kinect SDK. This file takes about 4-6 mb/min, depending on the time the face has been detected by Kinect. Each registry of this file contains 1504 values. First attributes of three rows of the file look like:

```
1352806825295,2012-11-13
12:40:25.295,4,7,6,12,200,1268,245,198,82,89,…
1352806825357,2012-11-13
```

As aforementioned, this information was not used in the research work reported in this Thesis, but is included here for completeness and to support future work.

4.6.4. Keyboard interactions

The file generated by the keylogger/mouse tracker app containing the keyboard interactions is a csv file with the following information:

- Time of the event: hour, minute, second and millisecond
- Type of event: ‘p’ for press or ‘r’ for release
- ASCII code of the key
- Representation of the key

Here is shown a log extract to see the fields generated:

```
11:14:10:343;p;55;7
11:14:10:390;r;55;7
11:14:10:875;p;13;RETURN
11:14:10:906;r;13;RETURN
```

As the interactions are exported as plain text, the files generated are not extremely huge. The size of the logs generated depends on the number of interactions performed during the session. A file with 2908 events registered (in a 40 minute session) takes only 60 kb.

4.6.5. Mouse interactions

The keylogger/mouse tracker app also generated another csv file with the mouse interactions. In this case, the file contained the following information:

- Time of the event: hour, minute, second and millisecond
- Type of event: ‘mv’ for movement, ‘prl’ for left button pressing, ‘rll’ for left button releasing, ‘prr’ for right button pressing and ‘rlr’ for right button releasing.
- X coordinate: coordinate X in pixels starting from the left part of the screen where the event has been registered
- Y coordinate: coordinate Y in pixels starting from the top part of the screen where the event has been registered

The event log generated looks like this:

```
11:04:34:718;mv;829;374
11:04:34:796;mv;829;375
11:04:34:796;mv;829;376
11:04:34:796;prL;829;376
11:04:34:859;rlL;829;376
```

These files usually need less space in hard disk (depending this on the interactions performed during the session). A file containing 21590 interactions (in a 50 minute session) takes just 503 kb of disk space.

4.6.6. Physiological signals

After recording the physiological signals, the Physiolab software offers the possibility of exporting the session data in two different formats: csv or excel. When exporting the data as a csv format, the timestamps captured during the session were not exported, so the files were exported as MS Excel files. As reported in section 4.5.1.c, there are two frequencies the data is generated by depending on the data source, the measured signals are recorded every 100ms, while other heart-related automatically generated indicators are calculated every 500ms. When exporting the data, not all the columns in a given row contain data from the same time, so before each signal column, another time column is exported indicating the time the following signal corresponds to. The following columns are exported in the excel file:

- Event: in case a timestamp has been taken during this registry recording, the name of the timestamp will be added here.
- HR_ (followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’): Heart rate, calculated number of heart beats in a minute. Its value is calculated each second, but the values are shown every 100ms, so, in the dataset, a row with a new value of HR is followed by 10 rows with the same value, until the next heart rate calculation has been performed
- SC A_(followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’): Skin Conductance, measured every 100ms.
- BPM_(followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’): Beats per minute, measured every 100ms
- RESP B_(followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’): Respiratory volume, measured every 100ms.
- IBI_(followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’):Inter Beat interval, the calculated time between two following heart beats. It is also calculated every second (like the heart rate, showing a value every 100ms), and can be obtained from heart rate, being: HR_=(60000/IBI_)
- TEMP A_(followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’): Body Temperature, measured every 100ms.

- HRV30_(followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’): Heart Rate variability averaged over 30 seconds, contains the difference between the maximum and minimum heart rate values in a 30s time window. This value is calculated with different frequency and shown every 100ms.

The following columns are calculated every 500ms and shown every 500ms (regardless if the data the row they are being displayed in has the same time than they). These columns are calculated from the Discrete Frequency Transform (DFT). The DFT is measured in a scale of 0 -.4 Hz. All the peaks at different frequencies in this indicator represent the power of different rhythms present in the inter-beat interval (IBI) measurement.

- HF_(followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’): High Frequency (.15 - .4 Hz)
- LF_(followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’): Low Frequency (.05 - .15 Hz)
- VLF(followed by columns ‘Time’, ‘Min’ , ‘Max’ and ‘S.D.’): Very Low Frequency (0 - .05 Hz).

The excel file contains then, rows with different times as can be seen in the following extract from the log:

```
00:00.100;69.64553833;00:00.100;4.34561157;00:00.100;6.82274246;(…);00:00.500;0.24969187;00:00.500;0.74066383;00:00.500;2.07531500;
00:00.200;72.96139526;00:00.200;4.34199047;00:00.200;6.82274246;(…);00:01.000;0.24969187;00:01.000;0.74066383;00:01.000;2.07531500;
00:00.300;72.96139526;00:00.300;4.33895922;00:00.300;6.82274246;(…)00:01.500;0.24962862;00:01.500;0.73357540;00:01.500;2.09714723;
```

In Figure 18 it can be seen which values are recorded at 10Hz and which ones at 2Hz. The time values of the signals with different frequency are not matched, so the generated csv file is erroneously built.

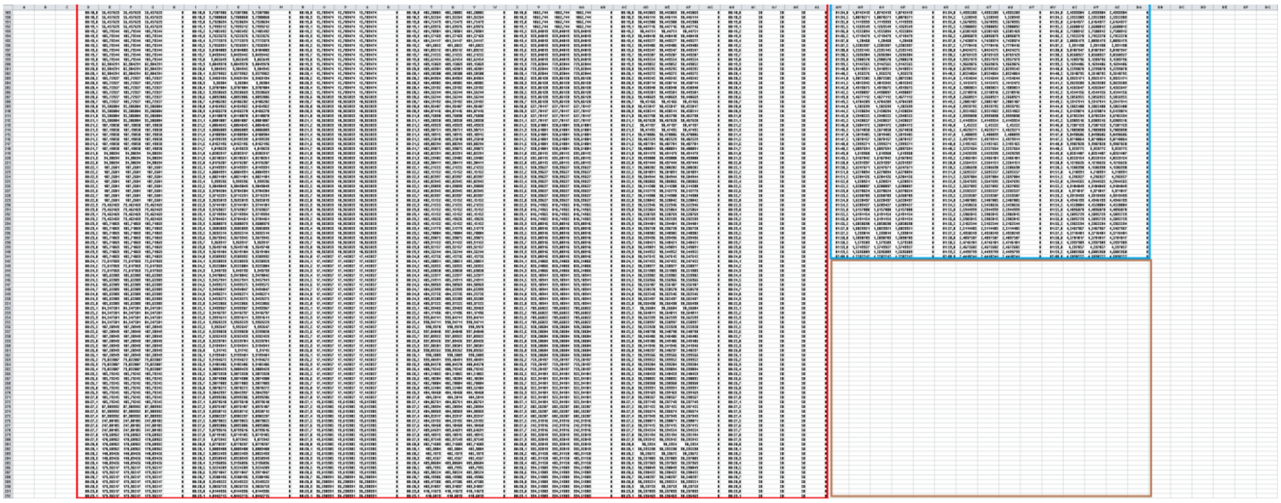


Figure 18. csv file generated by the Physiolab software. The red square includes values recorded at 10Hz, the blue square include values calculated at 2Hz. As they are sorted regardless the time, the columns including the values calculated at 2Hz are shorter than the others, leaving a blank spaces in 4/5 of the rows (brown space).

4.6.7. Screen recording

The result from the screen recording was a video file (.avi format) with a 1024x768 resolution (but stand 3, which had a different monitor with a 1152x864 resolution) at 50 frames per second. The videos take around 1-2mb/min.

4.7. Data preparation

Before performing data mining, the data collected has to be prepared as the data to be used needs to be cleaned, grouped and synchronized [187]. Since in this Thesis we do not consider the information from the webcam video, the Kinect video and the Kinect facial points, their preparation is not reported here. So we focus on the keyboard, mouse and physiological signals. In addition, we comment here on the indicators considered from the questionnaires on personality traits and report how the emotional reports were labeled.

4.7.1. Keyboard interactions

For processing the keyboard interactions, the log was split into tasks according the timestamps taken by the tutor during the session. Once the events were joint by task, the following indicators were generated for each group of interactions:

- Average time between two following key press events
- Average time per stroke (defined as the press of a key and its release)
- Number of key press events,
- Number of times a given key has been pressed
  - Backspace
  - Navigation arrows,
  - Delete
  - Tab
- Number of times a set of keys has been pressed
- Alphabetical characters

Other indicators were extracted from [78] based on similar criteria applied over combinations of 2 and 3 keystrokes (called digraphs and trigraphs). These indicators are:

- 2G_1D2D: The duration between 1st and 2nd down keys of the digraphs.
- 2G_1Dur: The duration of the 1st key of the digraphs.
- 2G_1KeyLat: Duration between 1st key up and next key down of the digraphs.
- 2G_2Dur: The duration of the 2nd key of the digraphs.
- 2G_Dur: The duration of the digraphs from 1st key down to last key up.
- 2G_NumEvents: The number of key events that were part of the graph.
- 3G_1D2D: The duration between 1st and 2nd down keys of the trigraphs.
- 3G_1Dur: The duration of the 1st key of the trigraphs.
- 3G_1KeyLat: Duration between 1st key up and next key down of trigraphs.
- 3G_2D2D: The duration between 2nd and 3rd down keys of the trigraphs.

- 3G_2Dur: The duration of the 2nd key of the trigraphs.
- 3G_2KeyLat: Duration between 2nd key up and next key down of trigraphs.
- 3G_3Dur: The duration of the third key of the trigraphs.
- 3G_Dur: The duration of the trigraphs from 1st key down to last key up.
- 3G_NumEvents: The number of key events that were part of the graph.

4.7.2. Mouse interactions

The steps taken to process the mouse interaction log were quite similar to those taken when processing the keyboard interaction logs. After splitting the logs into tasks, the following indicators were extracted from each group of interactions:

- Number of clicks: left button clicks, right button clicks and any button clicks.
- Distance the cursor has moved (Maximum, minimum, mean, standard deviation variation and number of events registered): to know how much the cursor has been moved.
- Speed cursor has been moved (Maximum, minimum, mean, standard deviation variation and number of events registered): to know how fast the cursor has been moved.
- Distance covered between events (Maximum, minimum, mean, standard deviation variation and number of events registered): The distance covered by the cursor between the following pairs of events:
    - Button press and the following button press events
    - Button press and release events
    - Button release and press events
    - Button release and the following button release events
- Euclidean distance between events (Maximum, minimum, mean, standard deviation variation and number of events registered): The Euclidean distance between the points where the following pairs of events happened:
    - Button press and the following button press events
    - Button press and release events
- Time between events (Maximum, minimum, mean, standard deviation variation and number of events registered): The time between the happening of the following pairs of events:
    - Button press and the following button press events
    - Button press and release events
    - Button release and press events
    - Button release and the following button release events
- Difference between the covered and the Euclidean distance between events (Maximum, minimum, mean, standard deviation variation and number of events registered): The Euclidean distance between the points where the following pairs of events happened:
    - Button press and the following button press events
- Button release and press events

4.7.3. Physiological signals

For preprocessing the physiological data, several steps were taken:

- Deleting the columns with a frequency lower than 10 Hz in order to have the same frequency in all our features.
- Deleting the duplicate columns. As all the variables in the dataset have right now the same frequency, is not needed the appearance of one time column per physiological signal, so we delete al but the first time columns. The columns with the minimum, maximum and standard deviation values are also removed, leaving this way only one observed value per physiological signal every 100 ms.
- The values of the remaining columns are split by task. Taking as reference the timestamps taken during the experiment. So this time we can ignore the values between tasks.
- Some temperature values are corrected. As on some sessions, the Physiolab software was configured to get the temperature in Fahrenheit degrees, and after that, in Celsius degrees, all the temperatures values were transformed to Celsius.
- The noise values were cleaned. To do this, two psychologists with a strong background in physiological sensing, provided a range where the values are supposed to be correct per physiological signal. If a value of a certain signal has a value outside of that range, it is considered noise and its value will be replaced by an interpolation of the previous values considered correct and the following values considered correct (see Figure 19).

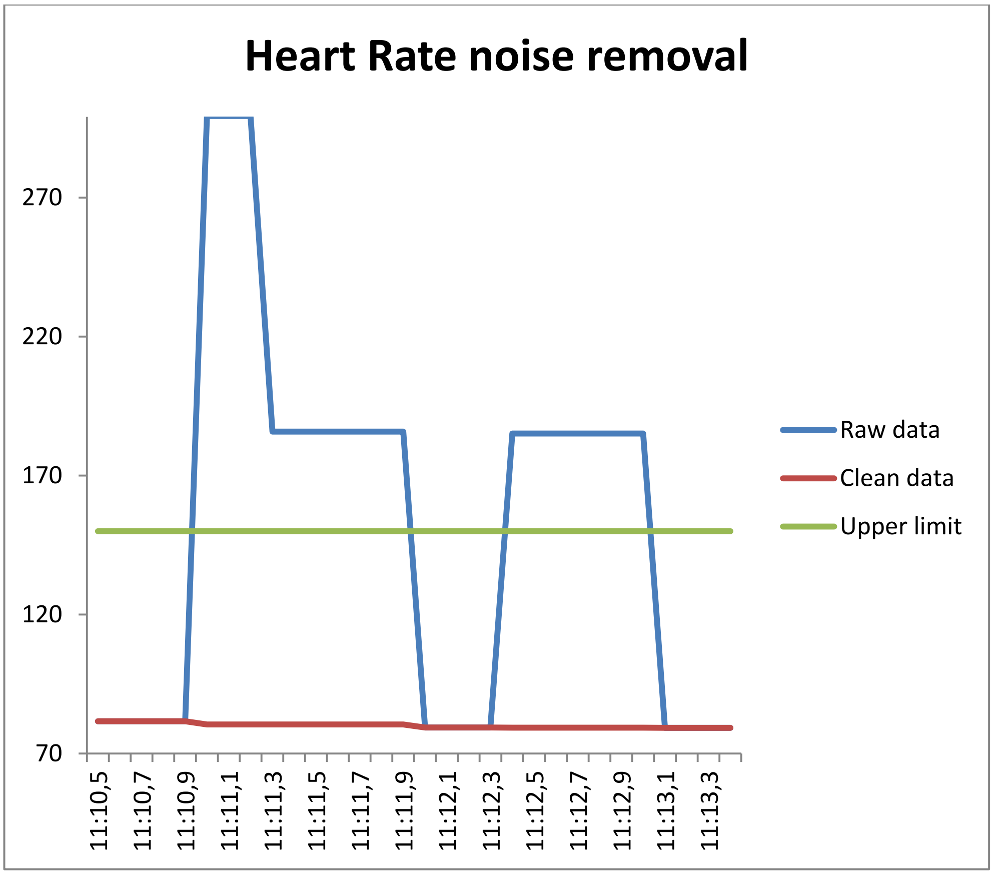


Figure 19. Heart rate noise removal, including the original signal (in blue), the upper rate threshold (in green) and the resulting clean signal (in brown).

- Once the values are clean, the mean value of the initial base line is calculated and subtracted from all the values in each task so the data can be normalized (see Figure 20and Figure 21) as done in [44,173].

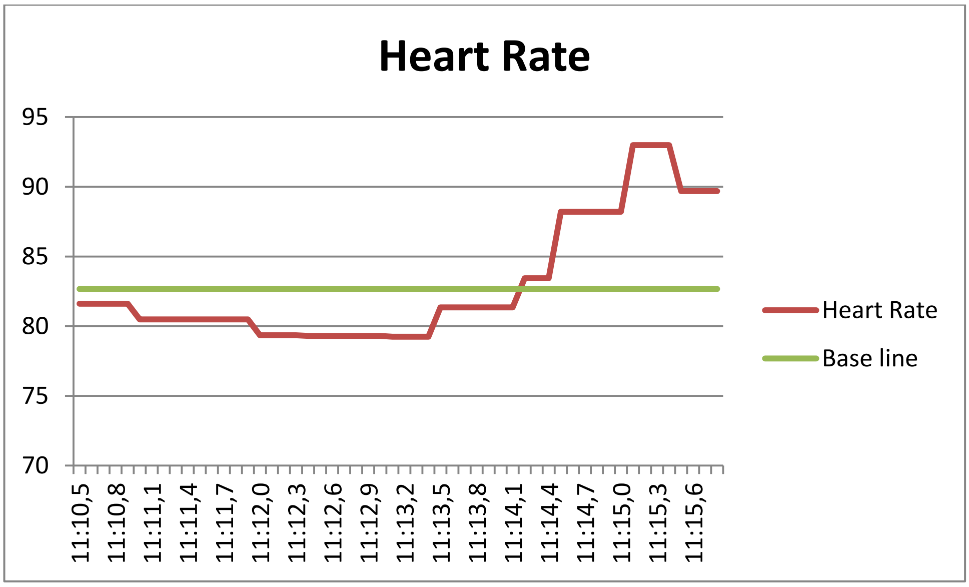


Figure 20. Heart rate and average heart rate taken as base line.

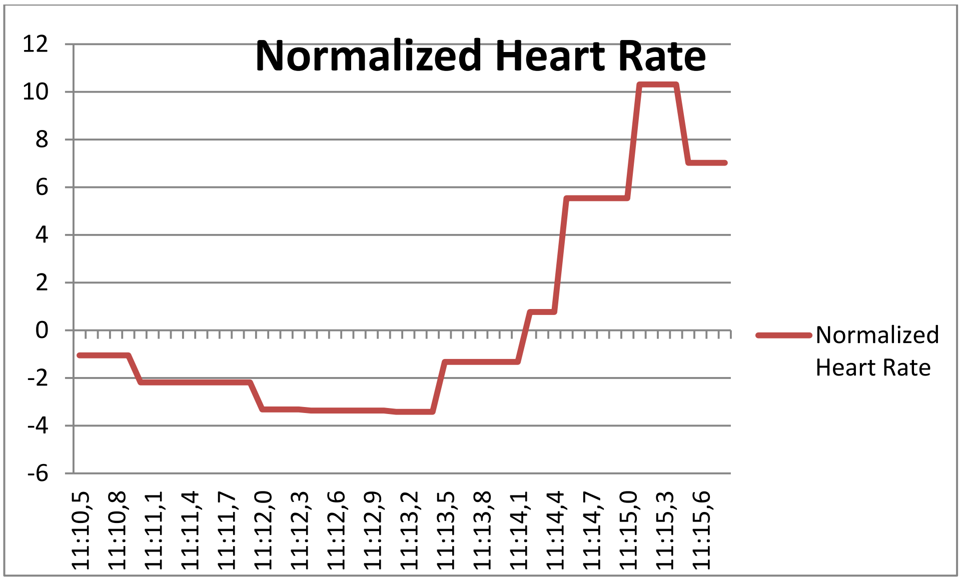


Figure 21. Heart rate values after being normalized.

- The last step taken is to group all the values for every task, generating for each task one registry and five columns: mean value, variance, standard deviation, maximum and minimum.

```
act2usr1ses10d09m11;_FROM_1_TO_fin-diffs.csv;1.997188524111982;164.82784544265743;12.838529722778128;60.89018557919039;(…)
act2usr1ses10d09m11;_FROM_2_TO_fin-diffs.csv;-0.8212037024438521;79.78923959258819;8.93248227496636;30.982554679190386; (…)
act2usr1ses10d09m11;_FROM_3_TO_fin-diffs.csv;1.450606879451832;126.52107696514486;11.248158825565403;60.88362427919037; (…)
act2usr1ses10d09m11;_FROM_4_TO_fin-diffs.csv;5.905527063877425;133.01967933964664;11.533415770691986;66.92608957919037; (…)
```

4.7.4. Questionnaires results

The questionnaires results were saved in a .csv file (average and standard deviations have been reported in section 4.4, in Table 5, Table 6 and Table 7), including the values for:

- Affective balance index (from the Positive and Negative Affect Schedule): the difference between Positive and Negative dimensions of the mood.
- Extraversion score (from Big Five Inventory): that provides information about the sociability of the participant.
- Neuroticism score (from Big Five Inventory): or tendency to experience negative emotions.
- Conscientiousness score (from Big Five Inventory): or tendency to be organized.
- Agreeableness score (from Big Five Inventory): related to sociability.
- Openness to experience score (from Big Five Inventory): Or curiosity for new things.
- General Self-Efficacy (GSE): the ability to face difficult situations.

4.7.5. Sentiment analysis

From the emotional reports, an automated indicator for the text valence was calculated. To do that, the MPQA subjectivity lexicon from University of Pittsburgh[10] was used. That lexicon provides a positive-neutral-negative labeling for each word, and the number of positive and negative terms in each emotional report was used also as an input data source. The score was calculated as follows:

$$\begin{aligned} &\mathit{Sentiment\ analysis\ score} \\ &\qquad = \mathit{Number\ of\ positive\ terms} \\ &\qquad - \mathit{Number\ of\ negative\ terms} \end{aligned} \tag{4.1}$$

It has to be noted that since the MPQA subjectivity lexicon is in English and our texts were in Spanish, Google translator was used, adapting the text to be translated term by term instead of being translated by sentences (trying to get an accurate translation of each term as the scoring will be by term counting). Nevertheless, as discussed in the future works section, this approach can be improved.

4.7.6. Emotional reports labeling

The texts from the emotional reports were labeled with two different criteria as follows:

- Two psychologists, with experience in motivational and emotional issues labeled each emotional report’s valence and arousal from 1 to 9 (following the SAM scale approach).
- An e-learning expert, with 10 years of experience in supporting learners in e-learning platforms, labeled each emotional report’s valence with one of the following values: positive, negative, neutral and positive-negative (i.e., when both positive and negative information was reflected).

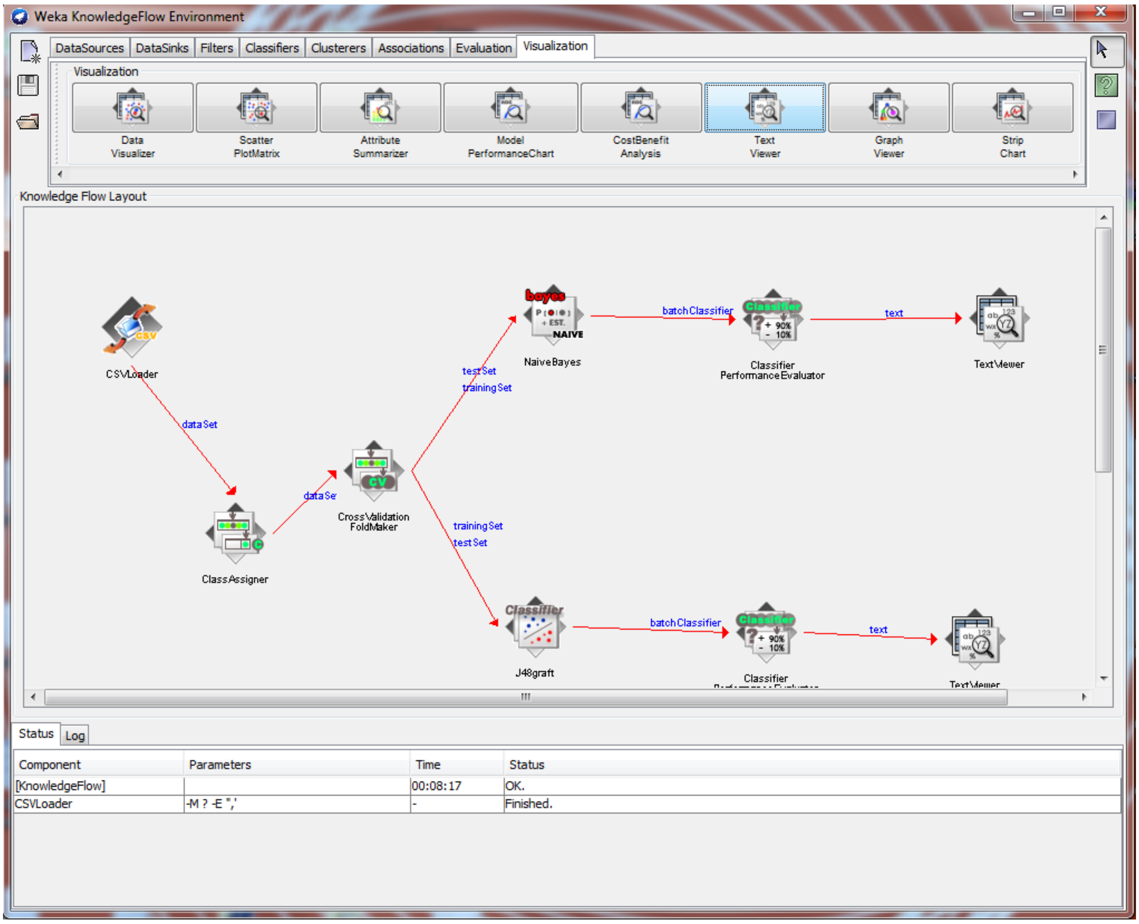
These labels were stored in different csv files, each one with its corresponding participant and emotional report indicator (i.e., 3 for the emotional report after the first Mathematical task, which was third in the list of tasks provided to the learner in the platform interface as reported in Figure 15, 4 for the emotional report after the second Mathematical task and 5 for the emotional report after the third Mathematical task).

When using the numerical labels to perform data mining, they were grouped into 3 different categories: positive (6-9 SAM score), neutral (4-6 SAM score) and negative (1-4 SAM score) as suggested elsewhere [22]. This way, the data can be handed in an easier way, helping to be used more easily when using it to trigger some reactions from the learning platform.

[10] http://www.cs.pitt.edu/mpqa/

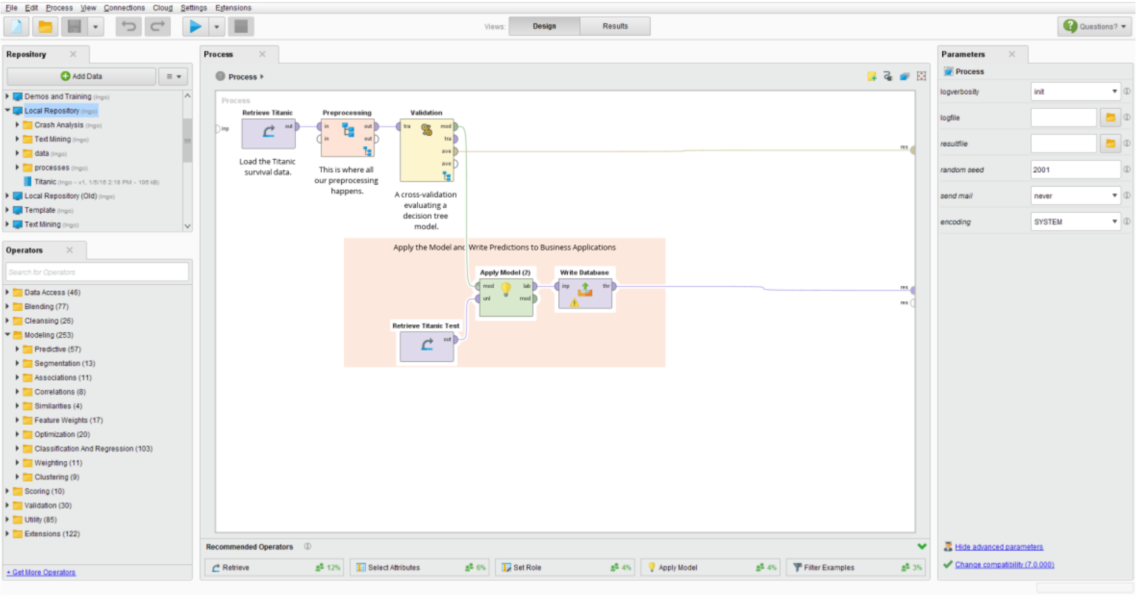
4.8. Data Processing

Once we have all the data ready to be processed, it was imported to a data mining tool. The tool used was Knime [25], an open platform to perform data mining. Knime is a visual tool (see Figure 24), based on Eclipse, which allows to create workflows by joining nodes that transform the data, similar to Weka's KnowledgeFlow [94] (see Figure 22) or RapidMiner [103] (see Figure 23).



Component	Parameters	Time	Status
[KnowledgeFlow]		00:08:17	OK.
CSVLoader	-M ? -E ",'	-	Finished.

Figure 22. Weka's KnowledgeFlow interface (screenshot downloaded from Weka's webpage)



Load the Titanic survival data.

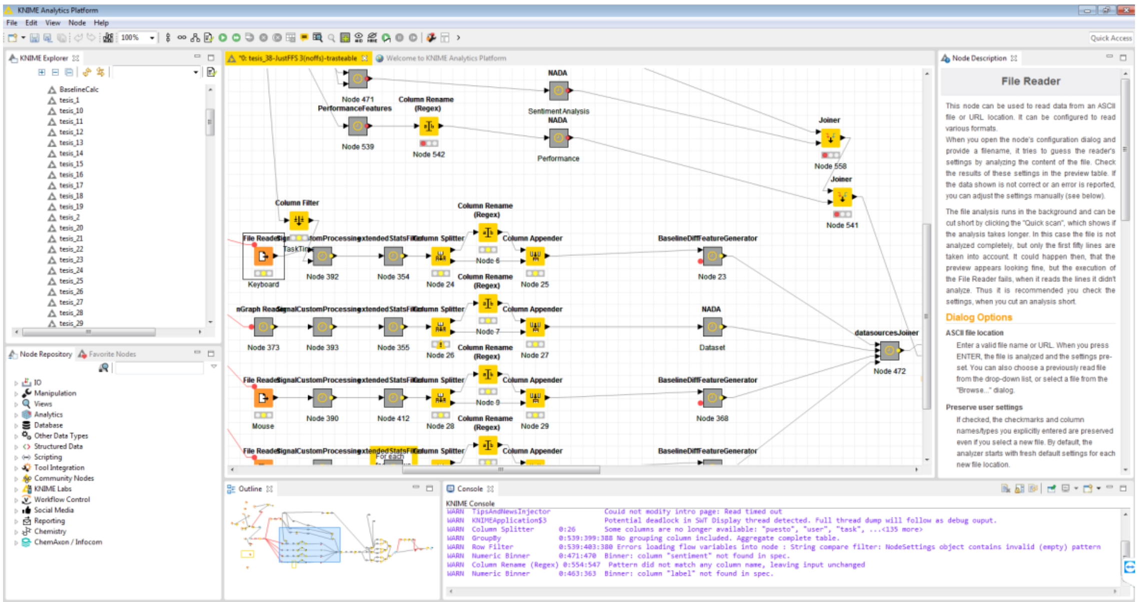
This is where all our preprocessing happens.

A cross-validation evaluating a decision tree model.

Apply the Model and Write Predictions to Business Applications

logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Figure 23. RapidMiner interface (screenshot from RapidMiner's webpage)



This node can be used to read data from an ASCII file or URL location. It can be configured to read various formats.

When you open the node's configuration dialog and provide a filename, it tries to guess the reader's settings by analyzing the content of the file. Check the results of these settings in the preview table. If the data shown is not correct or an error is reported, you can adjust the settings manually (see below).

The file analysis runs in the background and can be cut short by clicking the "Quick scan", which shows if the analysis takes longer. In this case the file is not analyzed completely, but only the first fifty lines are taken into account. It could happen then, that the preview appears looking fine, but the execution of the File Reader fails, when it reads the lines it didn't analyze. Thus it is recommended you check the settings, when you cut an analysis short.

Dialog Options

ASCII file location

Enter a valid file name or URL. When you press ENTER, the file is analyzed and the settings pre-set. You can also choose a previously read file from the drop-down list, or select a file from the "Browse..." dialog.

Preserve user settings

If checked, the checkmarks and column names/types you explicitly entered are preserved even if you select a new file. By default, the analyzer starts with fresh default settings for each new file location.

```
KNIME Console
WARN  TipsAndNewsInjector                 Could not modify intro page: Read timed out
WARN  KNIMEApplication$3                  Potential deadlock in SWT Display thread detected. Full thread dump will follow as debug ouput.
WARN  Column Splitter      0:26           Some columns are no longer available: "puesto", "user", "task", ...<135 more>
WARN  GroupBy              0:539:399:388 No grouping column included. Aggregate complete table.
WARN  Row Filter           0:539:403:380 Errors loading flow variables into node : String compare filter: NodeSettings object contains invalid (empty) pattern
WARN  Numeric Binner       0:471:470  Binner: column "sentiment" not found in spec.
WARN  Column Rename (Regex) 0:554:547  Pattern did not match any column name, leaving input unchanged
WARN  Numeric Binner       0:463:363  Binner: column "label" not found in spec.
```

Figure 24. Knime interface screenshot with a sample of the workflow implemented for this Thesis

The Knime tool was chosen as it provides a friendly and dynamic user interface, allowing the user to test and create different data mining computations very quickly and includes a high number of nodes, which allows not only to use well-known data mining algorithms but also pre-process the data, visualize it and import and export it. There is also the option to develop and download new nodes, which increases the potential of the Knime tool. Some of the most common Knime node packages include nodes with the Weka implementation of many data mining algorithms, nodes that allow developing our work with R scripts, nodes that include new ways of visualization (e.g. maps for geolocated data) or nodes to work with time series and time labeled data. Another important point to outline is the presence of a community of active users that interact in some online platforms and provide support in case of some problems.

The first step to be taken in Knime, as the data we have is stored in separate csv files, is to import them and join them in order to create a whole dataset with all the information to use as input for the different algorithms to use. A node for importing each file is added, and some preprocessing is made. As we are interested on detecting positive and negative values, we filter out the neutral values (as others have proposed [78]), focusing this way on differencing the states we are interested on detect.

In this step it could be seen that some labeling was not carried out properly as there were many registries from several data sources with names that did not match to the names of participant’s registries from other data sources. The consequence of this is a big dropout of registries as, when joining them by the combination of attributes participant and task, some combinations remain unmatched after joining two different tables, so the unmatched registries were filtered out.

When combining them together, the data could not be split in a problem-level granularity as the timestamps were taken at the beginning of every task (but not at the beginning of each problem). As the resultant registries reflected the whole data collected in every task, the emotional reports (collected at the end of each task) were used as a way to represent the overall emotion of each task.

When filtering all the registries and combining all the columns, the table contained around 150 rows and more than 500 columns. This table has too many columns compared to the number of rows it has. In fact, in data mining, the desirable situation is the opposite one, having a table with a bigger number of rows than columns. This is due to the possibility to generate overfitted models with tables that has a high number of columns. This is a common problem in data mining when is hard to get data to analyze, and is usually called “Curse of dimensionality” due to the high dimensionality.

To deal with that, there are some steps to take. Before applying the data mining algorithms, some columns have to be filtered out, but the selection of which columns to filter out cannot be done randomly. If discarded a column which provides a lot of information, the prediction results may be affected. First of all, columns with a low variance have been removed, as columns containing similar values in all their registries do not offer many information and the algorithms usually do not use them. By doing this, many columns are removed (as the mouse right button was not used by participants during the experiment, a lot of mouse right button related indicators are removed, something similar happens with pressing some arrow keys, etc.).

After that, the correlation between all the columns was computed, so this way we can see which columns are “similar”. Having highly correlated columns is not recommended in data mining as they offer “similar” information and mean a consumption of resources (memory and time when processing the model). This is why removing correlated attributes is a commonly used technique when removing attributes from data mining datasets. As for each mouse or keyboard interaction or physiological signal recorded, the recorded values were grouped into different indicators (mean, max, min, etc.) some of them are highly correlated, and most of them are discarded by the correlation matrix (shown in Figure 25), which was configured with a correlation coefficient threshold of 0.75. The selected column when different columns are found to have a high correlation is that with the highest number of correlated columns. Columns with higher correlation are darker in color. Red means negative correlation and blue means positive correlations.

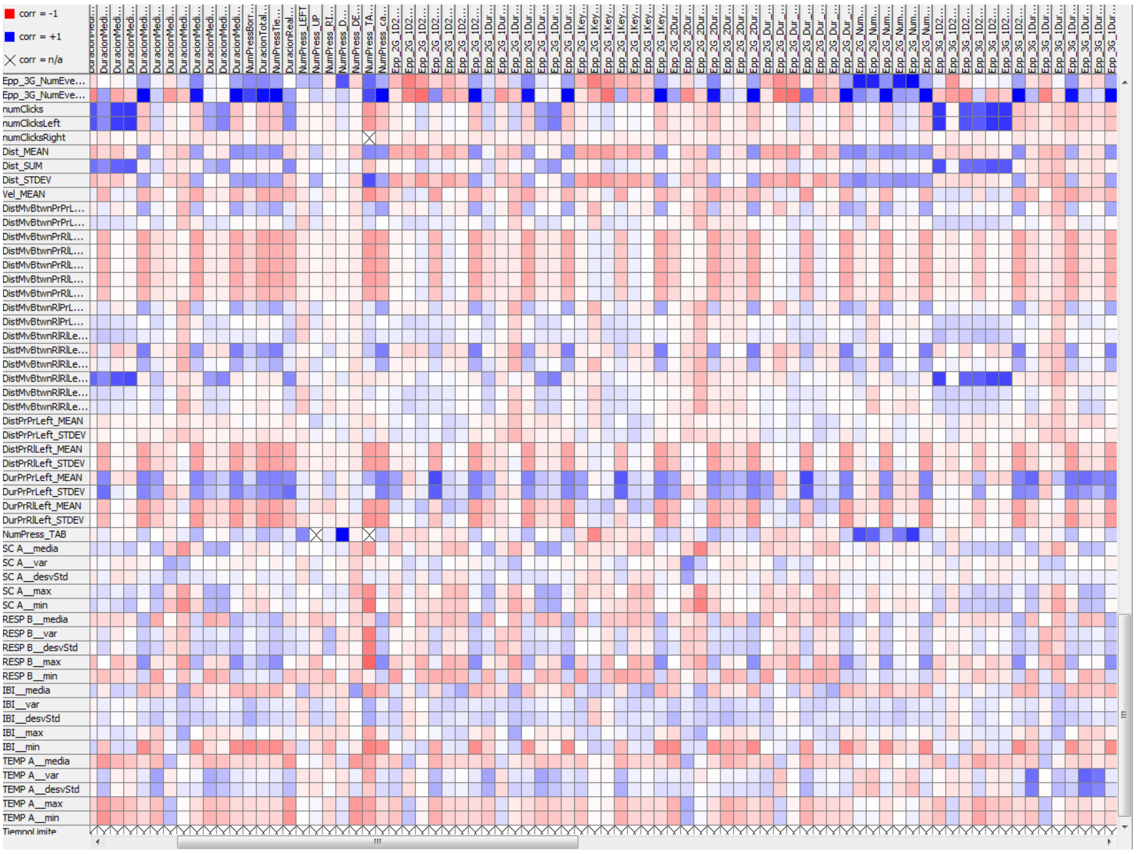


Figure 25. Correlation matrix visualization generated by Knime

After this, all the possible combinations of data sources are performed in the dataset for each one of the labels to be predicted. The labeling approaches that define the values to be predicted are the following:

- Valence given by the expert, with 10 years of experience in supporting learners in e-learning platforms.
- Valence given by two psychologists, with experience in motivational and emotional issues.
- Arousal given by two psychologists, with experience in motivational and emotional issues.
- Mean SAM valence values given by participants during the problems in each task.
- Mean SAM arousal values given by participants during the problems in each task.
- Average of the valence labels presented in the points 2 and 4 in this list.
- Average of the arousal labels presented in the points 3 and 5 in this list.

For each of these labels to be predicted, all the possible combinations (by joining the different data sources' attribute columns) of the data sources considered in Section 3 were generated (see Figure 26 for a graphical representation of them):

- Keyboard
- Mouse
- Sentiment Analysis
- Physiological signals
- Keyboard + Mouse
- Keyboard + Sentiment Analysis
- Keyboard + Physiological signals
- Mouse + Sentiment Analysis
- Mouse + Physiological signals
- Sentiment Analysis + Physiological signals
- Keyboard + Mouse + Sentiment Analysis
- Keyboard + Mouse + Physiological signals
- Keyboard + Sentiment Analysis + Physiological signals
- Mouse + Sentiment Analysis + Physiological signals
- Keyboard + Mouse + Sentiment Analysis + Physiological signals

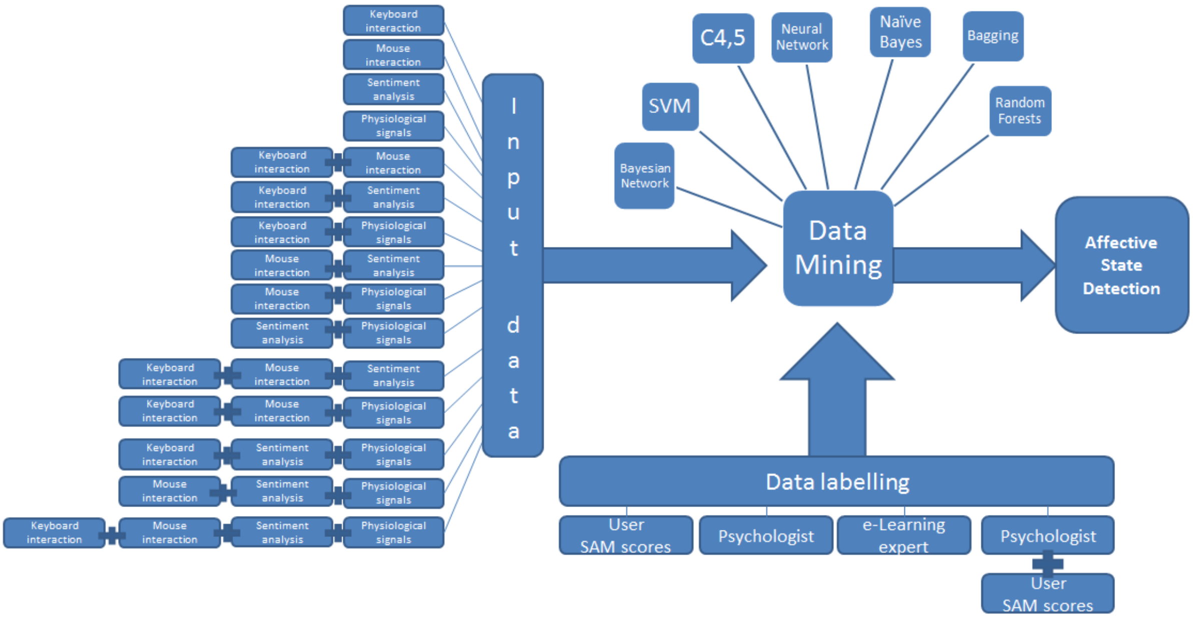

Keyboard interaction
Mouse interaction
Sentiment analysis
Physiological signals
Keyboard interaction + Mouse interaction
Keyboard interaction + Sentiment analysis
Keyboard interaction + Physiological signals
Mouse interaction + Sentiment analysis
Mouse interaction + Physiological signals
Sentiment analysis + Physiological signals
Keyboard interaction + Mouse interaction + Sentiment analysis
Keyboard interaction + Mouse interaction + Physiological signals
Keyboard interaction + Sentiment analysis + Physiological signals
Mouse interaction + Sentiment analysis + Physiological signals
Keyboard interaction + Mouse interaction + Sentiment analysis + Physiological signals
Input data
C4,5
Neural Network
Naïve Bayes
Bagging
SVM
Random Forests
Bayesian Network
Data Mining
Affective State Detection
Data labelling
User SAM scores
Psychologist
e-Learning expert
Psychologist + User SAM scores

Figure 26. Data sources combinations and labeling approaches followed

And for each one of these combinations, the supervised algorithms identified in literature (see section 2.2) commonly used for emotion detection were tested:

- J48: is the open source java implementation of the C4.5 algorithm by Quinlan, which builds trees based on the information entropy, being those attributes with a higher entropy closer to the root of the tree, and those with a lower entropy, closer to the leaves, which are the predicted value.

- Bootstrap aggregating (Bagging): is a machine learning meta-algorithm that generates new training sets from the original one, sampling with replacement from the original training set. This way variance and accuracy are slightly reduced and the number of training instances increases. The resulting model is obtained from averaging the output or voting, depending if the prediction value is numerical or not. In our case, the algorithm used to generate the different models was a Fast decision tree learner based on the information gain provided by each attribute (REPTree implementation in Weka).
- RandomForest: is an ensemble learning method based on the creation of different trees, each one with a different training set built from the original one by sampling with replacement. Each model also is generated from a subset of attributes. The predictions in the end are generated by voting. In this case, 10 trees were built for each model.
- Naive Bayes: is a technique based on the Bayes theorem assuming independence between the different features.
- Bayesian Network: starting from a Naïve Bayes approach, Bayesian networks allow to learn dependency and causality relations in the dataset.
- Support Vector Machines (SVM): Technique based on hyper planes that split the space depending on the class attribute, looking for the biggest distance between the closest instances to the hyper plane.
- Neural Network (NN): technique based on the combination of perceptrons (based on the behavior of a natural neuron), which calibrates the weights given to the different input variables depending on the output, trying to minimize the error.

Figure 26 depicts the three main methodological variables evaluated in this experiment (data sources, labeling approach and data mining algorithm used), and the different instances from each of those variables to be taken into account in the model generation. The datasets to generate the models will be generated form all the possible permutations from those variables. For all the predictions, cross validation was used, splitting the input dataset into 10 folds, using in each one of the 10 iteration the combination of 9 folds as training set and the remaining fold as test set.

4.9. Results

The results here presented show the accuracy of the prediction and the Cohen's kappa coefficient for the best combination of data sources and data mining algorithms used for each of the labeling approaches considered. The kappa coefficient was used as it takes account of agreement by chance between the predicted values and the observed ones, providing an reliable measure of model performance [79].

The accuracy shows the number of instances successfully classified from all the dataset. The accuracy is calculated as follows:

$$Accuracy = \frac{Number\ of\ instances\ correctly\ classified}{Total\ number\ of\ Instances} \tag{4.2}$$

The Cohen's Kappa coefficient indicates the inter-agreement between two observers (in this case, the observers are the reality, with the observed data and the prediction algorithms with the predicted data), taking into account the agreement occurring by chance. That is why Cohen's Kappa is commonly used for accuracy assessment to evaluate the behavior of the model. The Cohen's Kappa coefficient is calculated as follows:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{4.3}$$

Being Pr(a) the agreement between raters (accuracy in this case) and Pr(e) the probability of chance agreement. These are calculated as follows:

$$\Pr(a) = \frac{classified\ with\ the\ same\ value\ by\ both\ obervers}{Total\ number\ of\ Instances} \tag{4.4}$$

$$\Pr(e) = \sum_{i=1}^{Number\ of\ classes} \frac{Number\ of\ instances\ classified\ as\ i}{Total\ instances} \times \frac{Number\ of\ instances\ observed\ i}{Total\ instances} \tag{4.5}$$

735 different models were carried out since (as reported in the previous section, Figure 26) 7 labeling approaches, 15 data source combination and 7 data mining techniques were used as input in the data mining process (7*15*7=735). Therefore, to evaluate the results, an indicator was developed to rank the results combining the accuracy and the kappa score for the prediction results from each one of the labeling approaches. This indicator was a score calculated as follows:

$$ranking\ score = (1 + \kappa) \times Accuracy \tag{4.6}$$

The score was computed for every model tested, and, for each labeling approach, the best score for every possible data source combination was selected. For each one of the selected scores, the top three scores are reported in this section. For completeness, all the results are included in Appendix III (section 13.3).

The number of data instances may vary depending on the labeling approaches (due to the inconsistencies found in the identification of some data instances). When matching the label data with the data sources, depending on the participant's identifiers, some registries remain unmatched, so they cannot be used in the data mining process. The number of features considered also varies because as the registries may vary depending on the labeling approach, the values used to compute the correlation may slightly differ for each approach, so the filter may filter different columns in each labeling approach.

4.9.1. Results for labeling approach 1: Valence given by the e-Learning expert

This analysis considered 105 instances in the dataset, with the following class distribution: negative = 75 items; positive = 30 items.

The input features considered (37 in total) are compiled in Table 9.

Data Source	Number of Features
Keyboard	7
Mouse	18
Sentiment Analysis	1
Physiological	11

Table 9. Feature selection information for labeling approach 1

The top results obtained in the analysis are compiled in Table 10.

Score	Accuracy	Cohen's Kappa	Data sources	Algorithm
1,34363316	0,84761905	0,58518519	Sentiment Analysis	RandomForest
1,30204082	0,82857143	0,57142857	Mouse + Sentiment Analysis + Physiological	J48
1,27047619	0,82857143	0,53333333	keyboard + Mouse + Sentiment Analysis	Bagging

Table 10. Top results for labeling approach 1

As it can be seen, in this case the dataset has 105 rows to generate models from 37 attributes, obtaining from these, models that provide us accuracy values between 80% and 85%. It also has to be said that the best result for this approach was achieved from a single signal approach (sentiment analysis), being closely followed by two multimodal approaches.

4.9.2. Results for labeling approach 2: Valence given by two psychologist

This analysis considered 41 instances in the dataset, with the following class distribution: negative = 29 items; positive = 12 items.

The input features considered (22 in total) are compiled in Table 11.

Data Source	Features
Keyboard	9
Mouse	7
Sentiment Analysis	1

Data Source	Features
Physiological	5

Table 11. Feature selection information for labeling approach 2

The top results obtained in the analysis are compiled in Table 12.

Score	Accuracy	Cohen's Kappa	Data sources	Algorithm
1,372922169	0,85365854	0,60828025	Keyboard + Sentiment Analysis	RandomForest
1,269011217	0,82926829	0,53027823	Keyboard + Mouse + Sentiment Analysis + Physiological	NaiveBayes
1,246058249	0,82926829	0,50259965	Sentiment Analysis + Physiological	Bagging

Table 12. Top results for labeling approach 2

In this case, we only had 41 instances on the dataset, to generate models from 22 attributes. The accuracy values obtained from this labeling approach are quite high, and the Cohen's Kappa values are also fine. The top 3 scores in this approach are results from processing combinations of different data sources.

4.9.3. Results for labeling approach 3: Arousal given by two psychologist

This analysis considered 57 instances in the dataset, with the following class distribution: negative = 5 items; positive = 52 items.

The input features considered (30 in total) are compiled in Table 13.

Data Source	Features
Keyboard	9
Mouse	9
Sentiment Analysis	1
Physiological	11

Table 13. Feature selection information for labeling approach 3

The top results obtained in the analysis are compiled in Table 14.

Score	Accuracy	Cohen's Kappa	Data sources	Algorithm
1,13854557	0,9122807	0,24802111	Physiological	J48
1,07116383	0,89473684	0,1971831	Keyboard + Physiological	J48
1,01442825	0,87719298	0,1564482	Sentiment Analysis + Physiological	J48

Table 14. Top results for labeling approach 3

In this approach, we got very high accuracy rates but the Kappa values are very low, so the predictors are not performing as well as expected. This may be due to the class distributions of the dataset, where a 8,8% of the 57 rows in the dataset are negative and the 91,2% positive. In this case, the top 3 scores have been obtained by the same algorithm using always physiological data.

4.9.4. Results for labeling approach 4: Mean SAM valence values given by participants

This analysis considered 65 instances in the dataset, with the following class distribution: negative = 28 items; positive = 37 items.

The input features considered (34 in total) are compiled in Table 15.

Data Source	Features
Keyboard	8
Mouse	15
Sentiment Analysis	1
Physiological	10

Table 15. Feature selection information for labeling approach 4

The top results obtained in the analysis are compiled in Table 16.

Score	Accuracy	Cohen's Kappa	Data sources	Algorithm
0,85007692	0,66153846	0,285	Sentiment Analysis	RandomForest
0,82810779	0,64615385	0,28159539	Keyboard + Mouse + Sentiment Analysis + Physiological	SVM
0,79474529	0,63076923	0,25996205	Mouse + Sentiment Analysis + Physiological	SVM

Table 16. Top results for labeling approach 4

As we can see in this case, the accuracy levels are not quite high, and the Kappa values are low. The models generated from the 65 row and 34 column dataset do not offer great results.

4.9.5. Results for labeling approach 5: Mean SAM arousal values given by participants

This analysis considered 87 instances in the dataset, with the following class distribution: negative = 62 items; positive = 25 items.

The input features considered (33 in total) are compiled in Table 17.

Data Source	Features
Keyboard	8
Mouse	14

Data Source	Features
Sentiment Analysis	1
Physiological	10

Table 17. Feature selection information for labeling approach 5

The top results obtained in the analysis are compiled in Table 18.

Score	Accuracy	Cohen's Kappa	Data sources	Algorithm
0,88832067	0,74712644	0,18898305	Keyboard + Mouse + Sentiment Analysis + Physiological	Bagging
0,83732306	0,68965517	0,21411843	Mouse	RandomForest
0,8075979	0,72413793	0,11525424	Keyboard + Physiological	SVM

Table 18. Top results for labeling approach 5

In this case the highest scores offer us accuracy up to 75%, but the Kappa values are very low.

4.9.6. Results for labeling approach 6: Average of the valence values used in approaches 2 and 4

This analysis considered 47 instances in the dataset, with the following class distribution: negative = 24 items; positive = 23 items.

The input features considered (35 in total) are compiled in Table 19.

Data Source	Features
Keyboard	8
Mouse	16
Sentiment Analysis	1
Physiological	10

Table 19. Feature selection information for labeling approach 6

The top results obtained in the analysis are compiled in Table 20.

Score	Accuracy	Cohen's Kappa	Data sources	Algorithm
1,52185476	0,87234043	0,74456522	Keyboard + Mouse + Sentiment Analysis	J48
1,37696577	0,82978723	0,65942029	Keyboard + Sentiment Analysis	BayesNet
1,30695781	0,80851064	0,61650045	Keyboard + Sentiment Analysis + Physiological	BayesNet

Table 20. Top results for labeling approach 6

In this case, from the dataset containing 47 rows and up to 35 columns, the results obtained seem promising, with a top result with an accuracy of 87% and a 0.74 kappa. In this approach both keyboard and sentiment analysis appear in all the top models.

4.9.7. Results for labeling approach 7: Average of the arousal values used in approaches 2 and 4

This analysis considered 46 instances in the dataset, with the following class distribution: negative = 30 items; positive = 16 items.

The input features considered (37 in total) are compiled in Table 21.

Data Source	Features
Keyboard	6
Mouse	12
Sentiment Analysis	1
Physiological	11

Table 21. Feature selection information for labeling approach 7

The top results obtained in the analysis are compiled in Table 22.

Score	Accuracy	Cohen's Kappa	Data sources	Algorithm
1,24595055	0,80434783	0,54901961	Keyboard + Mouse + Sentiment Analysis + Physiological	RandomForest
1,20084492	0,7826087	0,53441296	Mouse + Sentiment Analysis + Physiological	J48
1,19021739	0,7826087	0,52083333	Keyboard + Mouse + Physiological	J48

Table 22. Top results for labeling approach 7

In this approach we get the best arousal prediction results. As both in the SAM score arousal and in the Psychologist scores arousal offered per results, when combining the scores from these approaches, the results obtained seem to be much better, increasing the kappa values and maintaining quite high accuracy rates.

4.9.8. Comparison between the labeling approaches

Once we have computed all the models, we can compare them in order to evaluate how the different labeling sources impact on the results obtained.

Score	Accuracy	Cohen's Kappa	Labeler	Target
1,52185476	0,87234043	0,74456522	SAM + psychologist	Valence
1,37696577	0,82978723	0,65942029	SAM + psychologist	Valence
1,37292217	0,85365854	0,60828025	Psychologist	Valence
1,34363316	0,84761905	0,58518519	E-learning expert	Valence
1,30695781	0,80851064	0,61650045	SAM + psychologist	Valence

Score	Accuracy	Cohen's Kappa	Labeler	Target
1,30204082	0,82857143	0,57142857	E-learning expert	Valence
1,27047619	0,82857143	0,53333333	E-learning expert	Valence
1,26901122	0,82926829	0,53027823	Psychologist	Valence
1,24605825	0,82926829	0,50259965	Psychologist	Valence
1,24595055	0,80434783	0,54901961	SAM + psychologist	Arousal
1,20084492	0,7826087	0,53441296	SAM + psychologist	Arousal
1,19021739	0,7826087	0,52083333	SAM + psychologist	Arousal
1,13854557	0,9122807	0,24802111	Psychologist	Arousal
1,07116383	0,89473684	0,1971831	Psychologist	Arousal
1,01442825	0,87719298	0,1564482	Psychologist	Arousal
0,88832067	0,74712644	0,18898305	SAM	Arousal
0,85007692	0,66153846	0,285	SAM	Valence
0,83732306	0,68965517	0,21411843	SAM	Arousal
0,82810779	0,64615385	0,28159539	SAM	Valence
0,8075979	0,72413793	0,11525424	SAM	Arousal
0,79474529	0,63076923	0,25996205	SAM	Valence

Table 23. Top 3 models from each data labeling approach, sorted by model score

As we can see in Table 23, valence models seem to provide better results than the arousal models. There, we can also see how the labeling approach providing best results is the combination of the SAM score with the labeling given by the psychologists. It is interesting as the labeling based exclusively on the SAM scores provided by the participants seem to provide the worst results from the top models analyzed.

4.9.9. Results Analysis

As it can be seen, sentiment analysis is present in 16 out of the 21 top scores, being the most used data source. It should be said that the 5 predictions where sentiment analysis is not present is in arousal predictions. This may make sense as the corpus used to label the terms from the emotional reports analyses the valence of the terms. Anyway, the obtained results show that the use of different data sources improves or equalizes the results provided by a single data source. This responds to the first hypothesis described in section 1.4 (**H1**), which drove the research conducted in this first stage: *Supervised data mining techniques on multimodal data sources improve the accuracy when detecting affective features to enrich learner modeling in task-independent educational contexts in comparison with single data sources.* In order to evaluate the validity of that hypothesis, the initial version of the AMO-ML methodology was designed and applied in the experiments carried out in 2012 (and reported, still unnamed in the following years [191,196,225]) to perform affective state detection has been defined and applied in an educational context-based experiment comparing the results of the predictions performed by data mining techniques using data from single data sources and combinations of data sources. As we can see in Table 25, combinations of data sources seem to provide the best results when performing affective state detection. To draw that conclusion, it has been needed to achieve the first objective described in section 1.4 (**O1.1**): *Evaluate different non-intrusive data sources to be used on emotion detection*.

Other point to take into account is the low kappa values obtained when predicting the arousal values. Both in the psychologists and in the participant given SAM scores the kappa values are below 0.3, but when combining these scores, the kappa increases. This may be due to the impact of the sentiment analysis, which provides a strong predictive capacity to the valence dimension.

The algorithm that appears the most in the top results is J48 (see Table 24), while the data source combination that appears the most is the combination of all the data sources (see Table 25).

Algorithm	Number of appearances
J48	7
Random Forest	5
Bagging	3
SVM	3
Bayesian Network	2
Naïve Bayes	1

Table 24. Algorithms in top results

Data sources	Number of appearances
Keyboard + Mouse + Sentiment Analysis + Physiological	4
Mouse + Sentiment Analysis + Physiological	3
Sentiment Analysis	2
Keyboard + Mouse + Sentiment Analysis	2
Keyboard + Sentiment Analysis	2
Sentiment Analysis + Physiological	2
Keyboard + Physiological	2
Physiological	1
Mouse	1
Keyboard + Sentiment Analysis + Physiological	1
Keyboard + Mouse + Physiological	1

Table 25. Data source combinations in top results

Regarding the labeling approach followed, we could see in Table 23 how the combination of the labeling generated by the psychologist with the participant-provided SAM scores offered the best model scores both in valence and arousal prediction. To arrive to this conclusion, the experiment was designed to address the second objective introduced in section 1.4 for this stage (**O1.2**): *Evaluate different emotion labeling approaches to be used as dependent variables.*

In this stage we have also aimed to deal with some of the research questions introduced in section 1.3:

- **Q1**: *Can the combination of different data sources in educational scenarios help to improve the affective state detection compared to single-data source approaches?*
    - From our results (Table 25) it seems that the combination of different data sources provides better results than using a single data source.
- **Q2**: *Which are the methodological aspects involved in the use and combination of different data sources with affective state detection purposes?*
    - In this sense, we have identified the methodological aspects involved in the use and combination of the proposed data sources, providing a detailed report in sections 4.2.1, 4.5.1, 4.6, 4.7 and 4.8.
- **Q3**: *Which affective state labeling strategies are more effective in real-world educational scenarios without penalizing aspects such as the intrusiveness of the approach proposed?*
    - From the results shown in Table 23, it seems that, the combination of the SAM and the psychologist labeling approaches is the one that provides best results in predicting both valence and arousal.

The results obtained could be improved if some aspects of the experimentation are refined. Nowadays, there is still a lot of work to do in the affective computing field (as the review of the state of the art reported in the literature show), and the current experiments are still building the basis for a strong affective state detection. In next section these aspects are discussed.

4.10. Discussion on Stage 1 results

After finishing the first research stage, some issues found during the experiment and data analysis have to be discussed. Here are the main issues to be taken into account in the next stage of this work:

4.10.1. Bad timestamp collecting design derived to too long time windows used

Due to the nature of the emotional field, the data instances to be analyzed should contain unique and exclusively the phenomena labeled, and as emotions are very short events, the desired scenario should be designed with very short tasks that strongly impact on the affective dimension of participants. In our case, instead of taking a timestamp every time a new problem was shown, the timestamps were only taken at the beginning and at the end of each set of problems, being this, a huge time window where many different emotions could have appeared. This may conduct to time windows were many emotions could be reflected, but only one class value (which may represent the last felt emotion, the strongest emotion of the task or whatever). That is why finally we decided to use only the data instances representing the emotional report tasks, as during the time the participant express his or her emotions, there is nothing that disturb the participant that may change the expressed emotions (but the mere act of expressing them, but that is something that will be always present when asking participants for feedback about their emotions), so it may be the closest we are to a “isolated” emotion.

4.10.2. Self-emotional report

It has been seen that the sentiment analysis is one of the most present data sources in the best models generated. That score that has provided so much affective information when detecting emotions (as seen in previous section) being calculated from affective oriented tasks, breaking the flow of the proposed mathematical tasks. In following experiments, that score should be calculated from task-related texts instead of making learners stop to type their emotions.

4.10.3. Physiological recording device limitations (data exported time marks and live data)

One desired functionality missing in the physiological recording hardware used is the streaming of the data recorded live. This way, the timestamps could be taken out of the physiological signal recording device and even the real time processing of the signals (that would be one of the final goals of the work here presented if a device like that would have been at hand for experimenting). If the raw signals are streamed, that would also help to merge all the different data sources and design the format of the data has to be exported, depending on the needs and not being “condemned” to export data with duplicated columns and in formats that need further processing to change its format. In the presented work, we used four J&J Engineering I-330-C2 systems, released in 2004, which captured noisy data, did not allow any data streaming as it used a closed software without providing any API and the export formats were really poor and slow. One of the biggest withdraws was also that the system recorded the signals in a time scale starting from the beginning of the experiment, instead of using the system clock, which would have helped to synchronize the data.

4.10.4. Timestamp synchronization

The previous issue is related to the synchrony of the data collected, which was faced by means of timestamps. This timestamp issue (i.e., taking the experiment timestamps from a device which does not use the same time reference than the other devices used) could have been solved automatically with a device that allowed to take automatic timestamps triggered by a signal (maybe a signal generated by the server every time a problem page is loaded), so that is important to know well the available devices to use and, in case the devices are going to be bought or developed, that functionality should be present to avoid human errors on timestamp taking. Due to the hardware used the synchronization of the signals was a duty that took much longer than it should. That is the reason why it is very important to have the right hardware to allow a correct and easy data capture (which is vital in data mining). To help the synchronization, the system time of all the computers used in the experiment was synchronized with the same time server, but the problem of the physiological signal device made that the timestamps were not exactly synchronized with the other signals. In this case, it was a "not so big" problem as the time windows we were using were huge (grouping all the values per task instead of per problem), but if a more detailed experimental design is to be done, that synchronization should be strictly taken into account.

4.10.5. Interaction devices usage in the tasks proposed

Regarding the design of the experience, it has also to be discussed the dependency on the mouse and keyboard behaviors to the task proposed. On the one hand, during emotional reports tasks, an intense use of the keyboard was needed, while on the other hand, in the problems tasks, using the mouse could almost be enough to solve all the problems (although keyboard was needed to provide the SAM scores).

Task and user interface also play a key role in the way the user interacts with the devices, so the indicators here presented depend on that. Some indicators have been proposed to be context independent (such as the different between the covered and Euclidean distance in mouse), but most of them may vary their values performing the same task with a different interface, and that should be taken into account when designing a user model based on interaction indicators.

4.10.6. Class attribute format and discretization

Other point to discuss is the limitations of the processing applied to the predicted values. As one of the experts providing the labels used a three-category format: positive, negative and neutral (really it was 4 categories as there were also some “positive-negative” registries). That was the chosen approach to follow when processing the labels, but, as mentioned when discussing the labeling format, the dimensional approach to define emotions can be “translated” in too many different ways of splitting the values into bins. The chosen approach was also elected due to its simplicity. The purpose of this research is to model the emotional dimension of learner so it can be used to offer an adaptive experience. Creating the positive and negative categories is a simple approach easy to handle when designing adaptive actions. But is this easiness one of the problems, as the less categories, the less adaptation to be offered (but the higher difficulty to detect the correct state). Here we can find that the more categories, the more complex the detection is and worse results obtained. Also, the detection of neutral states is hard to perform, but in this case we just wanted to evaluate how good can be data mining at detecting affective states discretized in different ways. That is another point of discussion, in a dimensional model as the adopted in here, where should be the threshold when dealing with states to be considered as neutral, positive or negative.

4.10.7. Inter-subject approach

It should be also discussed to what extent all the participants react the same way to the same situations. Data mining looks for patterns in big datasets containing the key attributes that generate the value to be predicted, but in affective computing, there are many works as have been seen in the state of the art aiming in many different directions, but most of them aiming to an inter-subject approach, assuming that there is a common pattern in affective behavior. To continue in this way, a huge amount of data is needed, from the most heterogeneous sample possible and, as we have seen in this work, it is not easy to collect this data. Nowadays, thanks to MOOCs, is “easy” to generate a course with thousands of learners, but the collection of data from them (especially the data proposed here) is still an intrusive issue (both at the physical and privacy levels).

4.10.8. Discussion summary

As we have seen, there are many points to discuss in this work. From the experimental design, that might be set out more fine-grained with a more detailed labeling on each problem instead of each task, with concrete problems and strongest emotion elicitation methods to the importance of the use of right tools that may ease the interoperability of the collected data. It also has been mentioned the importance of the context when working with interaction features. In this work, a way of splitting dimensional affective scores into categorical ones has been used, but many different alternatives can be proposed in this issue. At last, the discussion of the existence of common patterns of emotional behavior has raised the debate between an inter-subject approach and an intra-subject approach. All these issues are to be addressed in the second stage of this work. Some of the issues that have been discussed in this section include:

- In the experimental design, a system capable to take timestamps in all the important events of the experiment should be taken into account. That problem derived us to evaluate the affective states of the participants during sets of problems in contrast to evaluating the affective state for each problem.
- The good results provided by the self-emotional report proposed might be, in part, consequence of asking the participants to type about their emotions. That might be intrusive and in further stages the approach should move to perform sentiment analysis from the texts collected during the task (not asking them to type extra content).
- Related to the synchronization of the data (to be discussed in the following point), that is an important point when choosing the data collection devices to use. Other devices capable to ease the data synchronization (and less intrusive) should be used.
- The importance of a good data synchronization system. To do that it is needed an improved methodology collecting data and a system that takes into account the way every data source collects the data.
- Another important point when designing the experiment is to propose tasks where the data sources used are going to provide data. Very few keyboard interactions were collected during the tasks proposed so a sentiment analysis task was added in order to collect more keyboard data. In further stages we should propose a task where more keyboard interactions have to be performed.

- Regarding the labeling, there are a wide range of possibilities to evaluate. A small simplified labeling approach was used in this experiment, with positive and negative categories, but even in that approach, there are many open issues to evaluate (e.g. when collecting numerical data, how to discretize it to consider it positive or negative). Further research is required in next stages in that direction.
- We have followed a inter-subject approach in this work, but it would be interesting to perform an intra-subject experiment. Nevertheless, intra-subject experiments require a long-term design. For further stages, it would be interesting designing an intra-subject approach or see how to get closer to an intra-subject data processing from an inter-subject experiment in order to provide a more detailed learner model.

5. Transition Stage: Towards a Real World Learning Scenario

Once the first stage has finished, a second experimental iteration was planned. As the main goal in this thesis was developing a methodological and practical approach to perform affective state detection in real-world learning scenarios and evaluating how the different methodological aspects may impact the affective state detection, the results from that stage have to be carefully evaluated in order to design the approach to be followed in the second stage.

This stage has been designed as a transition stage between the stage 1 and the stage 2 as the initial version of the AMO-ML methodology described in the previous section is not fully deployed in this experiment (but some of the outcomes from the previous stage have been applied in this stage, as can be seen in section 5.1). This stage has been included due to its importance to the following stage, as its main goal takes place in a real-world learning scenario (as it is stated in the stage 2 hypothesis **H2** in section 1.4), in order to define a reference scenario to be used in the next stage of this research. In order to define that reference scenario, an experiment in collaboration with University of Valencia was carried out (described in section 5.2), where an ITS developed by them was going to be used in a real-world learning scenario, aiming to provide affective state detection capabilities to their system.

5.1. Lessons learnt from stage 1

At the end of the stage 1, we have detected some issues to take into account in future stages of the proposed research:

- The combination of all the proposed data sources has provided the best prediction rates.
    - In further stages, we will aim to include all the data sources evaluated in stage 1.
- Regarding the sentiment analysis data source, it seems to be the data source with a higher prevalence in the different sets of data sources providing best results, but it was evaluated from emotional texts, which may break the work flow of the learner.
    - It should be evaluated the possibility of performing sentiment analysis from texts extracted from the tasks to be performed by the participant, trying to avoid including affective-purposes tasks.

- Regarding the labeler, it has been seen how the different labeling approaches provide different results. One of the main problems found has been the resources needed to provide a labeling given by an external expert (temporal and human resources) as well as the potential problems that may arise from that kind of labeling (such as problems on the ids of the labels provided as seen in stage 1).
    - Although external labeling may provide better accuracy results, its inclusion in a real world affective state detection system can be discussed. In case a system with that goal aims to be used at large scale, it would be impossible to use that kind of labeling. In next stage, the use of self-labeling will be used as a main source of emotional labels.
- Regarding the tasks, the choice of the task has an impact on many methodological issues of the experimentation: the adaptation of the level of the task to the background of the participant, the use of some of the data sources may depend on the nature of the task, etc.
    - For the next stage, the design of a real-world task has to be one of the goals to follow, carrying also the experiment in a real world scenario where the system here proposed might be used (i.e. educational institution)
    - That task should also aim to the collection of as many data points as possible, requiring the use of the proposed data sources in a more intense manner than the one proposed in stage 1 (e.g. keyboard was rarely used until the emotional report in stage 1, so only a few keyboard interactions were collected in the tasks proposed in stage 1).

5.2. ITS Experiment

During the design of stage 2 and the end of stage 1, in the frame of the MAMIPEC project, some other experiments were carried out. Due to the collaboration held in the MAMIPEC Project with University of Valencia, a new experiment was held in year 2014. With this experiment, two different research lines aim to converge: the affective state detection system presented in this work and the ITS system developed by the MAMIPEC members of the University of Valencia [9]. The experiments carried out during the preparation of the second stage could be used to draw some methodological variables to analyze during that second stage, being used as pilot experiments.

5.2.1. Goals

The main goal of this experiment was the definition and evaluation of a reference scenario based in a real world-learning context to develop affective state detection experiments (in order to face the **H2** hypothesis in the stage 2). The idea behind the inclusion of this real world context is providing ecological validity to the AMO-ML methodology developed in this work. The next research stage will be built from the methodological conclusions obtained in this stage. This way we aim to hold an experiment that provides a new variable to the experiments to be performed, framing them into a real educational context with real students in their natural learning context. To this end, the following goals were set up:

- Evaluate the applicability of the methodological outcomes from stage 1 in a real world scenario (i.e. a real classroom).
- Evaluate the application of the approach followed in stage 1 using to other e-learning systems (different to the one used in stage 1).
- Evaluate the introduction of a new 2-stage detection approach, aiming to automatically detect and discard the neutral affective states and then evaluate the remaining data instances.
- Evaluate the use of a categorical labeling approach carried out by an external expert after the visualization of the recorded videos.

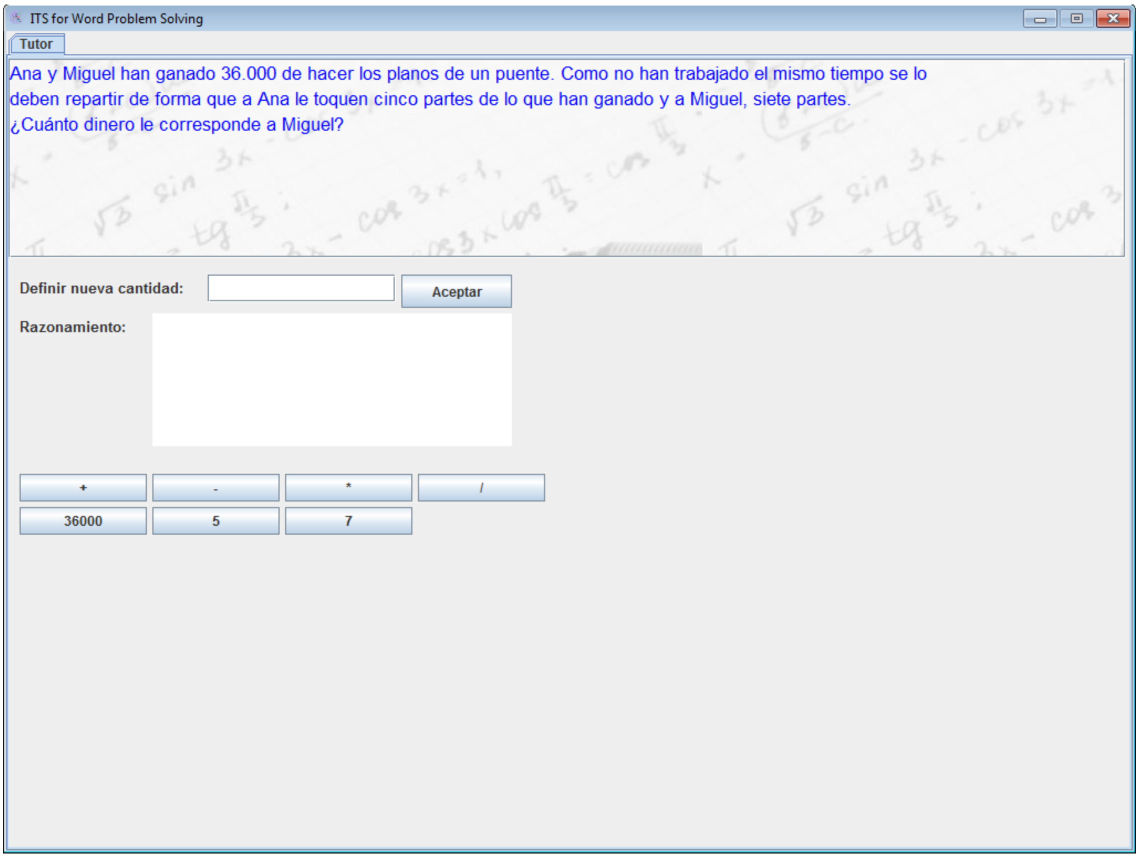
Due to the technical and temporal limitations of the proposed approach, two versions of the experiment were carried out at the same time:

- A simplified approach of the experiment, carried out with a total of 8 participants at a time, with the technical infrastructure available in the school (using only keyboard, mouse and webcam as data sources).
- A fine-grained version of the same experiment, which included physiological signals and Kinect. Due to technical limitations (as at this approach required more devices), only one computer was configured with this setup, so only one participant at a time could participate in the experiment (with a total of 2 participants following this approach). In this approach the participants also were required to perform a post-experiment evaluation of their reactions in order to enrich the labeling performed (which raised a temporal limitation in the experiment).

The data of the second approach followed was analyzed and the results reported in [192] are to be discussed here.

5.2.2. Context

MAMIPEC Project was proposed as a collaboration with University of Valencia. During the last years, they have been working in an ITS focused on algebra problem learning. That ITS provides the tools needed to solve problems by means of defining the different variables that have to be cleared along the problem solving process [9].



Ana y Miguel han ganado 36.000 de hacer los planos de un puente. Como no han trabajado el mismo tiempo se lo deben repartir de forma que a Ana le toquen cinco partes de lo que han ganado y a Miguel, siete partes.
¿Cuánto dinero le corresponde a Miguel?

Figure 27. Graphical User Interface (GUI) of the ITS developed by University of Valencia

In addition, a collaboration with the school Virgen de Mirasierra was set up for this experiment. By this collaboration, the experiment was able to be carried out in a real-world learning scenario, with real students providing tasks

5.2.3. Participants

Although the experiment was held with 10 participants, only 2 participants (one male and one female) were included in the fine-grained version of the experiment (including physiological sensors and Kinect). These two participants were 14 year old students, of the school. All the participants’ parents agreed to sign an informed monitoring consent.

5.2.4. Task

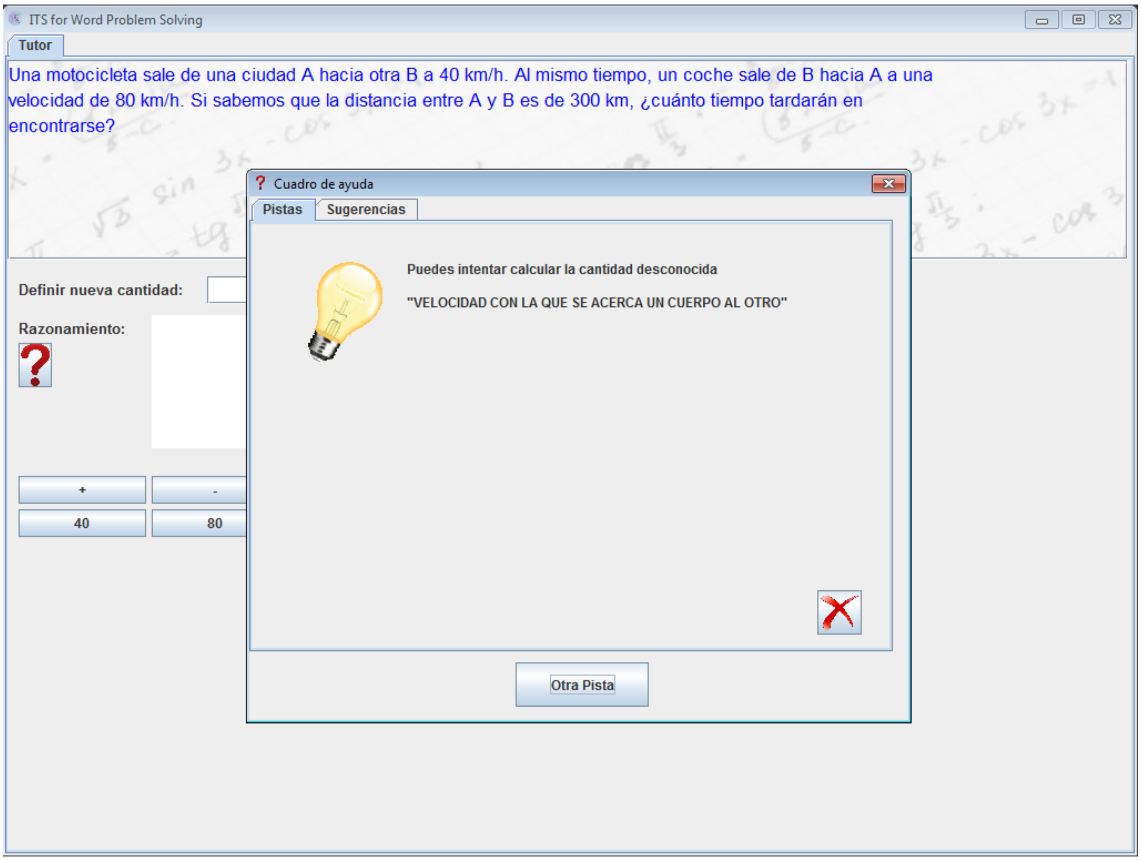
For this experiment, the subject chosen was, as in stage 1, Mathematics. As aforementioned, the tool used in this experiment for the task was different. An ITS developed by University of Valencia [7] was used in order to evaluate its use in a real-world context. The use of this tool introduces a several changes compared to the infrastructure used in stage 1. These changes are:

- The ITS used is a standalone tool, so it does not require the use of an internet browser nor internet connection.
- In contrast to the infrastructure used in stage 1, where participants only had the opportunity to provide the final result of each problem (without the possibility to annotate any amount during the problem solving), with this tool, students have

to define each amount or variable calculated from two previous amounts already defined until they define the final result of the problem proposed.

- The use of the ITS stablishes a limitation in the kind of problems that can be proposed, as this ITS only works with problems that can be solved in an arithmetic way.

All the participants were asked to solve the same problems. They also were provided the option of asking for hints during the third, fourth and sixth problems.



Una motocicleta sale de una ciudad A hacia otra B a 40 km/h. Al mismo tiempo, un coche sale de B hacia A a una velocidad de 80 km/h. Si sabemos que la distancia entre A y B es de 300 km, ¿cuánto tiempo tardarán en encontrarse?

Puedes intentar calcular la cantidad desconocida

"VELOCIDAD CON LA QUE SE ACERCA UN CUERPO AL OTRO"

Figure 28. Hint being shown during the problem solving. In the background it can be seen the question mark button (only available in certain problems) to request a hint

The proposed math problems can be found in Appendix II (in section 13.2.8).

5.2.5. Design

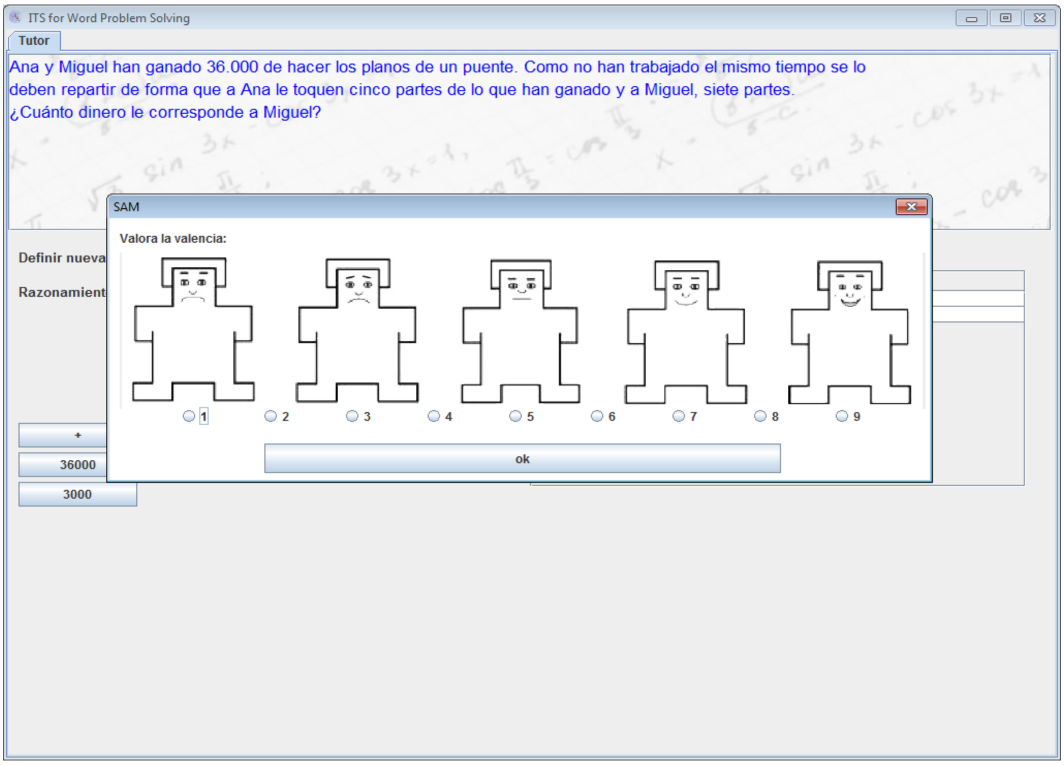
Assuming a multi-modal detection approach introduced in stage 1, a follow up objective of our research is to be able to detect emotions in a real world context (again with the support of three psychologists[11]). Besides major difficulties related to the particularities of real educational contexts (i.e., emotions are spontaneous and usually have a low intensity), this is a computationally challenging problem from a classification perspective, because of the traditionally high dimensionality of the input data.

[11] Mar Saneiro, Pilar Quirós and Raúl Cabestrero

To reduce the computational burden, in this experiment, a two-stage detection approach is proposed. At the first step, a two-class classifier to detect relevant time slots is proposed; and only relevant slots are then analyzed by a second classifier at a second stage. In this experiment, we focus on the first step, proposing a classification approach for filtering spontaneous and low intensity emotions in educational contexts. Second step first attempts have also been performed.

The proposal of this stage lies on an experiment carried out using an Intelligent Tutoring System (ITS) that focuses on teaching the resolution of story problems in an arithmetic way [9,12]. To account for the personality and physiological influence of the individual at expressing emotions [18], exhaustive data from two students was gathered using a similar data gathering approach that the one described in the stage 1. In addition, and motivated from the initial version of the AMO-ML methodology drawn in stage 1, some adaptations were included in the ITS by its creators in the University of Valencia, in order to be used in this experiment with affective purposes [11]:

- The inclusion of the Self-Assessment Manikin scale at the end of each problem solved was added.



Ana y Miguel han ganado 36.000 de hacer los planos de un puente. Como no han trabajado el mismo tiempo se lo deben repartir de forma que a Ana le toquen cinco partes de lo que han ganado y a Miguel, siete partes.
¿Cuánto dinero le corresponde a Miguel?

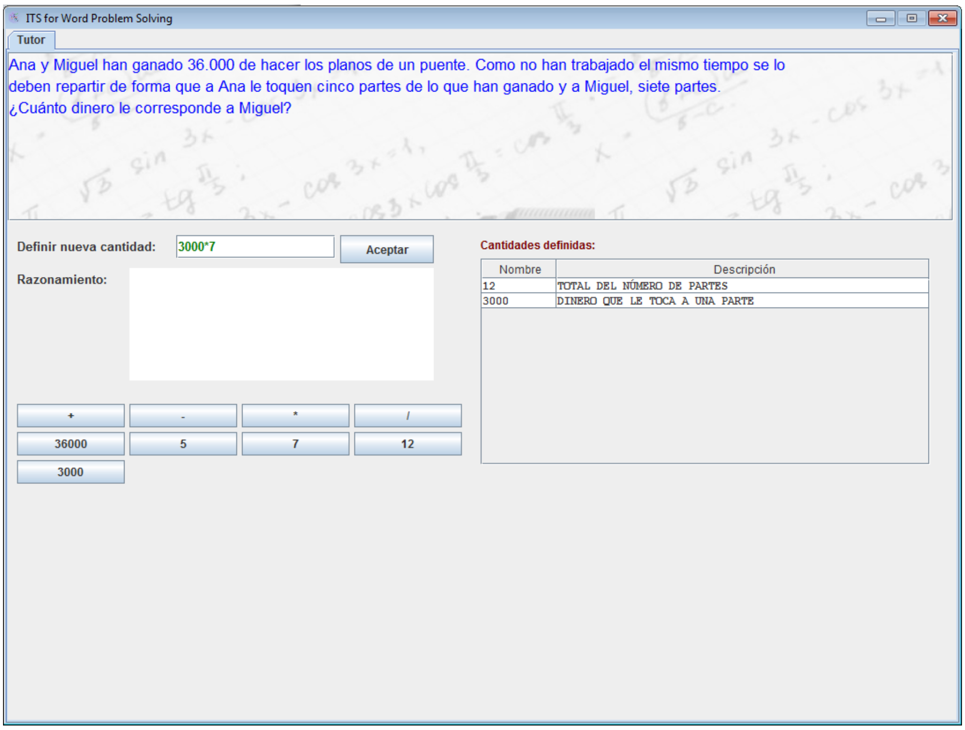
SAM

Valora la valencia:

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9

Figure 29. Self-Assessment Manikin implementation shown during the experiment

- In order to collect keyboard interactions, when defining a new variable, the participants had to type a short text explaining the new variable defined.



Ana y Miguel han ganado 36.000 de hacer los planos de un puente. Como no han trabajado el mismo tiempo se lo deben repartir de forma que a Ana le toquen cinco partes de lo que han ganado y a Miguel, siete partes. ¿Cuánto dinero le corresponde a Miguel?

Definir nueva cantidad: 3000*7

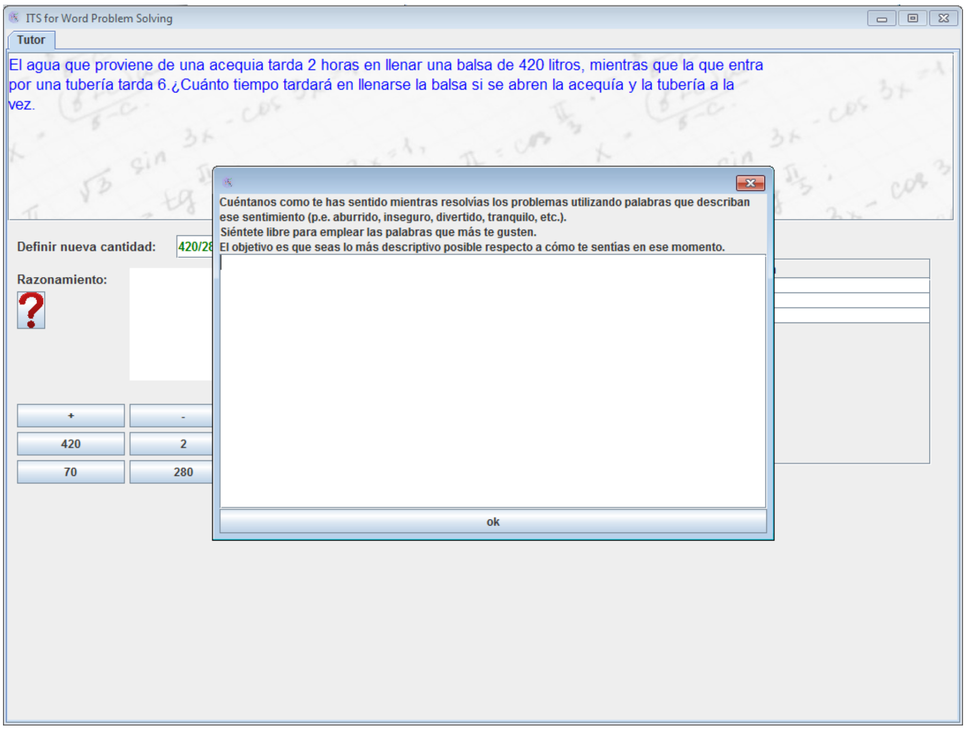
Razonamiento:

Cantidades definidas:

Nombre	Descripción
12	TOTAL DEL NÚMERO DE PARTES
3000	DINERO QUE LE TOCA A UNA PARTE

Figure 30. ITS interface showing the "Razonamiento" (reasoning) text area for the participant to type her reasoning when solving a new variable of the problem

- At the end of the experiment, a text area was shown to the participants asking them to type down their feelings during the experiment.



El agua que proviene de una acequia tarda 2 horas en llenar una balsa de 420 litros, mientras que la que entra por una tubería tarda 6.¿Cuánto tiempo tardará en llenarse la balsa si se abren la acequía y la tubería a la vez.

Definir nueva cantidad: 420/28

Razonamiento:

Cuéntanos como te has sentido mientras resolvías los problemas utilizando palabras que describan ese sentimiento (p.e. aburrido, inseguro, divertido, tranquilo, etc.).
Siéntete libre para emplear las palabras que más te gusten.
El objetivo es que seas lo más descriptivo posible respecto a cómo te sentías en ese momento.

Figure 31. Final form asking for the participant to type down her emotions during the experiment

Data collection was done by running an experiment in a real Mathematical class of 14-year old students that were asked to solve a series of 6 mathematical story problems[12] adapted to their age and knowledge, using a modified version of the ITS presented in [12]. This ITS was modified to capture emotional data (through self-reporting) at several stages:

- Before the student starts solving any problem, she had to fill the Attributional Achievement Motivation Scale [145] to explain the causes of the academic achievement.
- After completing each problem, the student had to report on her affective state (valence and activation) by using the Self-Assessment Manikin (SAM) scale [39] (as done in stage 1).
- At the end of the series, the student wrote a descriptive self-report detailing aspects related to her affective state during problem solving that she considered relevant (also, as done in stage 1).
- Once the experiment finished, the participant was invited to visualize the experiment recording with a psychologist who had followed the experiment remotely.

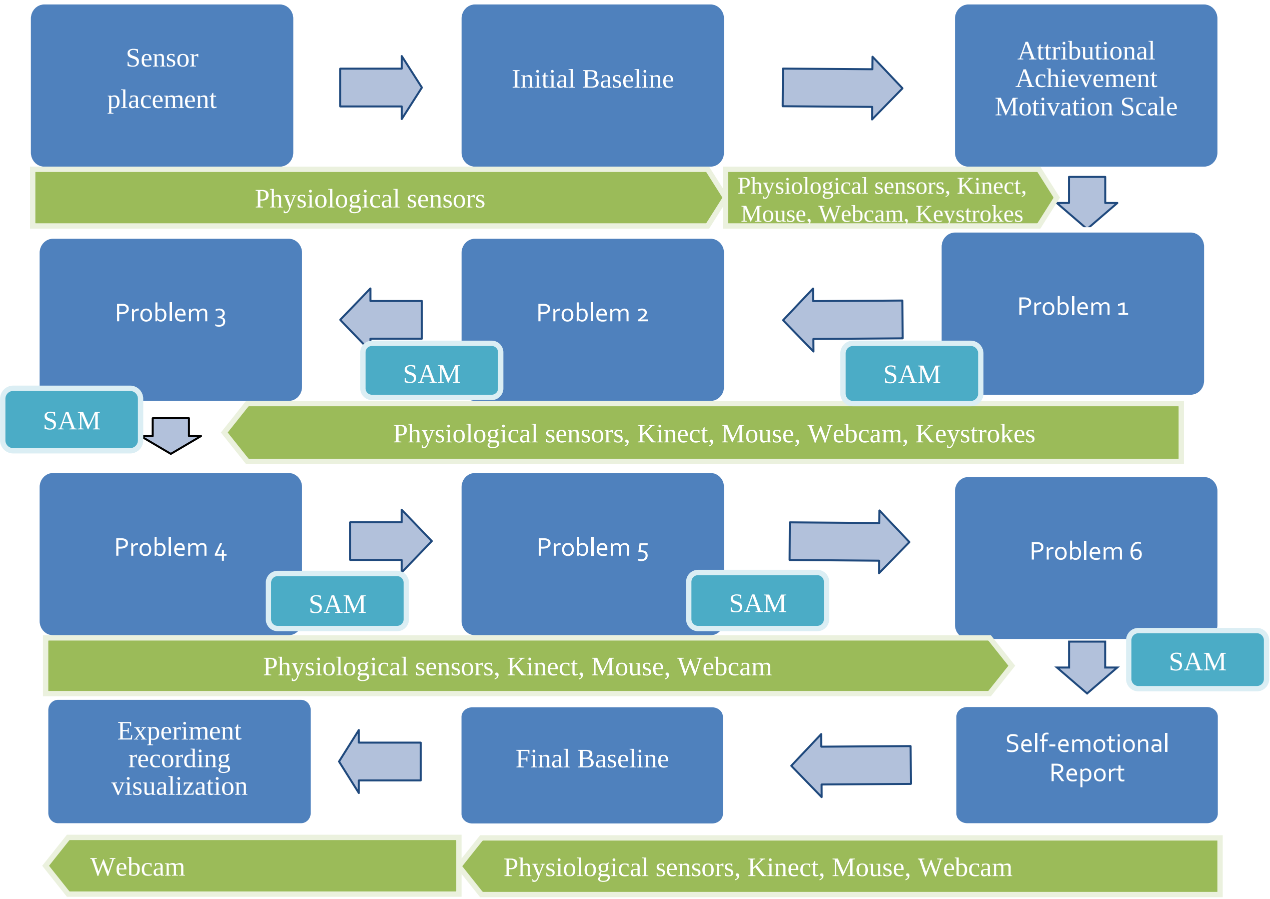

Sensor placement
Initial Baseline
Attributional Achievement Motivation Scale
Physiological sensors
Physiological sensors, Kinect, Mouse, Webcam, Keystrokes
Problem 3
Problem 2
Problem 1
SAM
SAM
SAM
Physiological sensors, Kinect, Mouse, Webcam, Keystrokes
Problem 4
Problem 5
Problem 6
SAM
SAM
SAM
Physiological sensors, Kinect, Mouse, Webcam
Experiment recording visualization
Final Baseline
Self-emotional Report
Webcam
Physiological sensors, Kinect, Mouse, Webcam

Figure 32. Experimental structure, including tasks and data to be colleted

[12] Selected by psychologists Raúl Cabestrero and Pilar Quirós

5.2.6. Data recorded

During the session, and other than the self-reports, exhaustive data from two students were gathered by using the following input sources:

- Physiological data: following a similar setup to the one used in stage 1, heart rate, breath volume, skin conductance and temperature captured at a frequency of 10 Hz with a J&J Engineering I-330-C2 system in a single comma-separated (.csv) file, each row representing 100ms of the experiment (the finest granularity of all the logs recorded). To have the baseline of the physiological signals for each user, a 3-minute recording while she was asked to stay relaxed were taken at the beginning and at the end of the experiment session.
- Interaction data (e.g., problem being solve, hint requests, correct and incorrect user actions, etc.) of events reported by the ITS were stored in a .csv file.
- Video data (webcam video and a desktop recording) stored in a single .camrec file generated by Camtasia Studio that contains a synchronized recording of both data flows. To focus on the emotional analysis without cumbersome video processing, webcam videos were analyzed by a human expert who reported in a .csv file both, movements performed by participant (including body part and type of movement) and emotions observed using the methodology described below.

To allow for the corresponding synchronization, the first column of all files corresponds to the timestamp of the event collected.

5.2.7. Data labeling

As we have seen in section 2.3, labeling emotions is one of the most controversial and critical open points in emotion detection as the way it is done may suppose some limitations for the future processing. Video data were watched by a psycho-educational expert trained on emotions detection who applied the methodology proposed in a previous research [204] to detect the facial expressions and body movements associated to emotions elicited while solving the ITS problems. For this, the expert simultaneously analyzed the webcam video (with participants' face) and the corresponding desktop recording (with the learning tasks carried out). The annotation process followed a mixed (judgment and sign based) approach and used the predefined tags from the previous research, enriched by adding the "movement duration" feature to consider the length of each movement. In this experiment, we are using a categorical and a dimensional approach. On the one hand, after each problem, participants were asked to report their affective state regarding the dimensions of valence and arousal using the SAM (dimensional approach). On the other hand, emotions tagged by the expert follow the categorical approach.

In contrast to the approach followed in stage 1, and taking advantage of to the low number of participants evaluated (2 participants), an external expert[13] viewed all the

[13] A psychologist: Mar Saneiro

video recordings, labeling the emotion of the participant following a categorical approach. The labeling performed in this experiment was a highly detailed labeling as the external expert provided labels of all the emotions detected during the experiment, annotating also the exact moment of the beginning and end of the external manifestation of the emotion.

Here, after coding the emotions when watching the recorded videos, she compared her notes with emotion aloud elicitation process carried out with the participants just after the experiment. Each participant was played the recordings with her face and desktop and asked to spontaneously comment aloud how she felt during the experiment (emotion aloud approach). She was also asked by the expert when she detected some movement or expression of relevance uncommented by the participant. With this information, the expert assigned relevant time-stamped educationally emotional labels when appropriate to each recording. Labels used in this experiment followed the EmotionML [220] and explicitly include emotions that could specifically appear in a learning context such as anxiety, confused, concentrated, frustrated, happy, shame or surprise, as well as none (absence of emotion).

5.2.8. Data preparation

First step to take once the data has been collected, and taking into account that the labeling has been provided in a continuous way, with no time windows predefined, is identifying a temporal window in order to split the data into chunks to be evaluated by the algorithms. As we could see in section 2.3.3 there are different approaches in the way to define the time window split criteria. In this case, we followed the criteria showed in section 2.3.3.a, using a fixed time window. In order to identify an appropriate for the length of the temporal window used to analyze the physiological variables, a recursive analysis was performed by MAMIPEC project collaborators[14] with different time windows (1 min, 30 sec, 20 sec). First of all, each signal was studied separately from raw data (sampling rate of 100 ms). The initial baseline was disregarded, because, i) some signals such as temperature and skin conductance did not reach stabilization until almost 10 minutes after the beginning of the recording and, ii) much of this phase is revealing reactions to the experimental situation. Therefore, we ended up using the final baseline as a baseline indicator of no reaction, in particular, the last 20 second time window before the end of the baseline.

With the previous scope in mind, those MAMIPEC project collaborators proceeded to average raw data into the aforementioned temporal windows, to identify (looking for significant differences between the final baseline and the task using ANOVA and its corresponding post hoc comparisons) which of them could reveal a better compromise between a sufficient level of results granularity and significant discrimination capacity of the signal changes triggered by the performance on the ongoing task. Temporal windows of 1 min and 30 sec were discarded due to the excessive smoothing of the signal that could be masking the small oscillations that tend to appear in such low

[14] Pilar Quirós and Raúl Cabestrero

intensity emotional reactions. Finally, the analysis ended up revealing significant changes from the final baseline for the 20 sec window that could be linked to the different phases of the learning task. Similarly, other authors have used this same temporal window to identify the presence of affective reactions in learning situations, reporting that it is even possible to detect several subjectively emotional states within this time window [178].

After determining the temporal window, we proceeded to truncate raw data into 20 second beans (200 values each) for each signal and every problem (initial and final seconds of each problem were disregarded to make sure that each window had 200 values). With these data, an ANOVA was conducted for each of the temporal windows of the problem and the last temporal window included in the final baseline, indicating which temporal windows (per subject and signal) were significantly different from the final baseline ($p < 0.001$). Those were labeled as “activated”.

As a result, in the preprocessing, rows were grouped in 20 second time slots, and a new feature vector per time slot was created, with the following contents: i) a sequential identifier for the time slot; ii) four new features, one for each physiological signal, indicating if the values of that signal during that time slot has suffered significant variations regarding the student baseline (binary: 0= no; 1= yes); iii) the sum of the previous four features to show how many signals suffered variations; iv) the number of incorrect actions carried out by the learner in the ITS during the time slot, v) the number of hints requested in the time slot; vi) a feature for each part of the body involved in a movement, containing the fraction of the time slot the part has been moving, and vii) a feature for each type of movement observed, containing the fraction of the time slot that type of movement has been occurring. Values assigned to these attributes, in the range [0, 1], depend on the movements reported by the expert.

This grouping operation yielded a total of 246 registries, each with 31 features.

5.2.9. Model generation results

The data gathered, labeled and preprocessed has been used in a typical classification setup as the one introduced in stage 1. Nevertheless, due to the labeling produced in this experiment, a new approach for data processing was used. A 2-step classification approach where initially neutral states are to be filtered out and then, those time windows left are evaluated in order to perform a finer grain classification of the affective state.

To filter spontaneous and low intensity emotions in educational contexts as corresponds to the first proposed stage aimed to detect relevant time slots, each data registry has been labeled with a binary value. Using the emotional labeling performed by the expert, 0 has been used if no emotion is present in the time slot, and 1 is some emotion has been detected. This labeling allows to adopt a classical binary classification setting to predict relevant time slots from an affective perspective (i.e., those where some emotion is detected). The SAM labeling was not used in this analysis as it was obtained at the end of each problem, thus would not make much sense to assign its value to all the 20secs registries reported per problem (problem resolution average length was 6 minutes).

To generate the model, we have used the J48 algorithm (Weka's Java implementation of the C4.5 algorithm [183]). We have tried other algorithms such as Naïve Bayes [110] and used Bagging [41] but we did not obtain a significant variation in the results. We have also tried a number of dimensionality reduction methods such as Backward Feature Elimination (BFE) [93] and Principal Component Analysis (PCA) [245], but they did not improve the results obtained either. The accuracy of the model produced has been assessed by using a leave-one-out cross validation on the labeled data. 74.8% of registries were correctly predicted when evaluating if there was an emotion or not in a given time slot.

For the second step (i.e., analyzing only relevant slots were emotions were detected) preliminary experiments to predict the specific emotion of each registry using this 2-step method yielded an accuracy rate of around 62.6%. When trying to predict the specific emotion from the initial dataset, without discarding previously automatically before the registries detected as non-emotional, the top result achieved offered an accuracy rate of 59.7%. These results point the proposed 2-step emotion detection approach as a promising way to simplify the emotion detection process. The results obtained in this section can be found in Appendix IV (section 13.4).

5.2.10. Conclusions

In order to advance some of the open issues in emotions detection in educational contexts, where emotions are spontaneous and tend to be of low intensity, we have proposed a two-stage detection approach that combines two classifiers, aimed to filter spontaneous and low intensity emotions from diverse emotional data sources gathered from educational contexts. The first one (binary) decides if there are emotions in a given time slot from the participants interaction or not. The second one predicts the emotion (in those slots that detected its existence). In this way, the emotions detection process focuses on the relevant time slots, improving accuracy and reducing processing time, especially with large datasets, making it more appropriate for real time processing.

Results from this research will be used to build a new version of the ITS used in this experiment that provides emotional formative feedback by replacing the current help-on-demand mechanism by a rule-based system that is able to use interaction data to both provide automatic recommendations and adapt the content of the messages, according to the user's affective state. The emotional support to be provided by the emotional formative feedback will be defined with the TORMES methodology [207] in terms of content (selecting and specifying the information provided within feedback), scheduling and timing (e.g., delayed vs. immediate feedback, feedback on work in progress vs. on complete work), sequencing (e.g., from general to specific) and presentation (e.g., multi-sensorial feedback delivery) as well as the learner characteristics involved, including cognitive and metacognitive issues when seeking help and feedback, considering proactive vs requested feedback and evaluating post-feedback behaviors, perception of feedback, acting upon feedback. In this way, we can research when is feedback effective, what kinds of feedback are effective, and whether there are individual differences in seeking and using feedback, as well as associated effects and outcomes, such as effects of feedback on current problem performance, next problem performance, transfer, retention, future learning, motivation, affect, achievement orientation.

5.3. **Methodological Outcomes**

The experiment described in this section was held in between the design stage of stage 2, and the end of stage 1. The main goal of this experiment was to perform an initial definition of a real-world learning scenario to apply the initial version of the AMO-ML methodology described in stage 1. Here is a list of the methodological points evaluated in this experiment and how they will have an impact on stage 2:

- Real-world scenario:

  The experiment has been held in a real world scenario (a real high-school) with real students performing tasks related to their current formation (related to the issue discussed in the previous point).
  The celebration of future experiments in a real-world context (in contrast with stage 1 experimentation) should be a requirement in order to evaluate the future applicability of the proposed approach in a real world scenario.

- 2-step prediction approach:

  In this experiment, a 2-step prediction approach was evaluated (described in section 5.2.9). In this approach, we aimed to evaluate for each time slot if the system considered presence or absence of affective state and then, taking into account only those time slots where an affective state was supposed to be, perform a finer-grain prediction aiming to detect which affective state was taking place in those time slots.
  In stage 2 we will aim to include this approach, comparing it to the approach followed in stage 1 (trying to detect the affective states directly in a 1-step prediction approach).

- New task proposed:

  In this experiment, the task proposed was different to the one proposed in stage 1. This change is due to the inclusion of the ITS system developed by the University of Valencia. Although this change was not very significant, as the subject chosen was still Mathematics, some changes were included due to the technological platform used (including the adaptations performed in the ITS for its use in an affective-oriented experiment, commented in section 5.2.4).
  This issue should be further explored in stage 2, as some points such as the adaptation of the task proposed to the target participants should be taken into account.

- Preprocessing techniques:

As mentioned in section 5.2.9, some preprocessing techniques were used in order to reduce the dimensionality of the dataset.
In stage 2 the impact of using those techniques on the results of the models generated has to be evaluated.

6. Stage 2: Practical and Methodological Development in Real World Scenario

After the definition of the AMO-ML methodology for affective state detection (using machine learning techniques with data collected from different data sources) carried out in stage 1 and the specification of the reference scenario based in a real-world learning context, this stage aims to combine the outcomes of those previous stages. Thereby, this stage's main contribution is to provide a new version of the initial version of the AMO-ML methodology for affective state detection by means of machine learning techniques in learning scenarios developed in stage 1. This new version aims to be applied in a real-world learning scenario (based on the reference scenario from the transition stage) as well as to evaluate further finer-grain methodological issues found during the previous stages. This will drive us not only to celebrate an experiment in a real classroom, but also to limit some methodological aspects previously evaluated in order to propose a realistic sustainable approach (i.e. it is not realistic suppose the labeling of an affective state detection system in production phase can be carried out by human labelers as it was done in stage 1).

In contrast with stage 1, the methodological variables to be analyzed in this section, remain, most of them, in the data preprocessing process, aiming to deal with the lack of information provided in this sense (as it was seen in section 2.2). These variables are, methodologically, more fine-grained than those that were object of study in stage 1 and have arisen from the experiments already carried out. Additionally, the focus is also to be set on some points form the experiment depicted in section 5 (transition stage), which have been included in the methodological issues faced in this stage (e.g. the 2-step classification approach).

It should also be pointed one of the main contributions from this stage: the interaction data normalization approach. This normalization proposes using a similar approach to the normalization performed on the physiological signals in stage 1 (depicted in section 4.7.3) over the interaction data (i.e. data collected from mouse and keyboard interactions). This proposal aims to get rid of some aspects such as the user skill (when using an interaction device) when building models from different subjects in an across-subject approach.

6.1. Goals

Stage 2 was designed following an incremental approach from the outcomes of the stage 1. The main goal of this stage is the application of the approach proposed in that first stage in a real-world scenario similar to the one defined in the transition stage. The application of the designed AMO-ML methodology in real-world conditions has driven many methodological decisions in the design of this stage 2, from the celebration of an experiment in a real-world context (as done in section 5.2) to avoiding external experts interaction (as the labeling process described in section 4.7.6 for stage 1 and in section 5.2.7 for the transition stage). By avoiding using experts, we aim to carry out a feasible approach on a large scale (where the use of experts supervising any aspects of all the potential users of a system like the proposed in this work is unaffordable), knowing that that is a methodological aspect that may impact negatively on the results.

This study also aims to get closer to a real-world scenario by means of using low-cost open hardware sensing devices in affect detection [215], in addition to keyboard and mouse, two non-intrusive information sources have been used regularly in affect state detection [76,120].

Other way to get closer to a real-world learning scenario is involving ordinary daily practices of learners (e.g., in learning subjects such as English as a Second Language), which in our case consist free text tasks. Because of this we are avoiding tasks that involve typing a fixed text several times, which have been employed in some previous studies [76,228]. However, affect detection from free text data [76] and realistic scenarios present lower accuracy results than those settings that use fixed text inducing affect through stories or video clips [122].

Additionally, other goals are to be achieved in this stage, going these other goals in two directions:

First, to use a normalization technique of keyboard and mouse interaction features by means of generating an initial user baseline. In that direction is set the main hypothesis to be evaluated in this stage (**H2**):

> *In real-world learning scenario based inter-subject experiments, the use of a reference state to normalize each user interaction related data provides more robust models when detecting affective features to enrich learner modeling in educational contexts.*

This aims to perform normalization similar to the one performed with the physiological signals (depicted in Figure 20 and Figure 21) to interaction-based data sources. Once this normalization approach and the addition of the required interaction baseline have been performed, its impact on the generated models should be evaluated (**O2.1**). The goal of this normalization is to evaluate intra-subject changes from data collected in an inter-subject experiment.

The second direction to be followed in this second stage of the research aims to evaluate some open methodological points found during the experimentation in previous stages. After finishing stage 1 experimentation, performing the first affective state models, many decisions taken were questioned. Many methodological points addressed in stage 1 (and in many related works) are addressed with a lack of comparison between alternatives, with the most common approaches being replicated with no justification in some points. Realizing that has driven this work to a second stage where, in contrast to the goals in stage 1, the issues evaluated address finer-grain methodological variables. There are a wide range of methodological points still to be discussed that are not going to be addressed in this section (but will be discussed in section 11 but two objectives have been set in order to evaluate the impact of some data preprocessing techniques (**O2.2**) during the model generation and different label discretization criteria (**O2.3**).

Other changes introduced from stage 1 include dealing with open issues such as time frequency to use in detecting emotional state changes. Here, the approach depends on the given task, and can be either a time window (e.g., a 10-minute window of keystroke interactions is used in [76,115]) or a number of events recorded (e.g., 600 events in [122]). We are also taking into account performance features related to the learning task because they may have an impact on the participant’s affective state as reported in [137,171]. In this sense, in stage 1 we used a group of tasks as a time window for detecting the affective state of the user. In this second stage, we aim to use a more detailed time window (i.e. perform a affective state detection per task).

Other issue related to the affective state detection that has to be evaluated is the affective model to use that represents those states to be predicted. Characterizing and labeling affective states to train data mining models has been a long-term issue in affective computing research and, in particular, in education modeling [178]. There has been extensive work on using experts’ knowledge to label users’ affective states [31,122], with increasing success even in the wild (e.g., students in a school) [32], but here we focus on getting a readily available model for detecting affect and thus facilitate a prompt reaction from the learner that can be used in situations where trained expert labeling is not available (e.g. at home) or a prompt reaction to an unexpected situation is needed. To this end students label their own affective state using well-known and widely used psychologically validated scales that represent the intensity of their affect reactions in two different dimensions, namely valence and arousal. Both valence, which represents the attractiveness/averseness of an affective state, and arousal, which refers to the level of activation of an affective state, have been extensively applied to collect participants’ affect [39,122].These two dimensions have been chosen as they account for most of the variance in affective reactions [39] and have shown significant correlations with performance [230]. This finding is relevant to our research, since cognitive demands have an impact on typing behavior [42]. In our case we are using both dimensions in daily tasks, as they provide a readily available information source for collecting data over time.

Keyboard, mouse and especially physiological signals and sentiment analysis have been extensively employed in affect detection studies using data mining techniques. However, there are still several problems related to this approach, such as the difficulty of comparing case studies because of the different emotional states and feature subsets considered and the different methodological decisions taken when generating the datasets [122].

In this stage we will follow the main steps that are commonly depicted in related works (and were also followed in stage 1 from the guidelines introduced in section 3.2): i) collecting and labeling data, ii) extracting relevant features (i.e., those providing plausible discriminant data for the given task being carried out by the subject), iii) training classifiers and iv) recognizing emotions [120]. As will be discussed below, throughout these tasks there are modeling options that can be further researched if they are taken as methodological variables. Given the lack of individual user data in educational contexts [188], another critical common problem is high dimensionality, that is, having many more features to describe users´ interactions than the number of available instances [31,122]. In this work, we have evaluated as a modeling variable the appropriateness of different preprocessing and reduction techniques to face this high dimensionality problem.

All these issues, and other related topics, will be discussed in this stage through a real-world case study based on detecting the affective state of the user from keyboard and mouse interactions as well as physiological signals and sentiment analysis. We are going to evaluate the modeling issues we have found in previous experiments that may affect prediction, thus showing that there are some benefits in further exploiting key issues involved, such as using a baseline model of the user's individual keystroke and mouse dynamics, preprocessing the dataset by applying class balancing and dimensionality reduction techniques as modeling variables, and adopting a simplified dimensional approach for labeling the user affective state.

6.2. Methodological variables

In this section we are going to describe the methodological variables to be analyzed in this stage to improve the initial version of the AMO-ML methodology. Some of these variables (included in section 6.2.1) are addressed from the previous stage discussion (section 4.10), while some other new variables are introduced (sections 6.2.2 to 6.2.7).

6.2.1. Methodological variables discussed in stage 1

At the end of stage 1, some issues remained open and were discussed in section 4.10. Here, the way those issues is going to be addressed in this section is going to be briefly described.

6.2.1.a. *Bad timestamp collecting design derived to too long time windows used*

One of the problems found in stage 1 is that the data registries used to generate the models were too long, using only the self-report data to feed the models. In this stage, timestamps are going to be collected at the end of each task, as well as a labeling. The tasks proposed in this stage take 210 seconds each one.

6.2.1.b. *Self-emotional report*

The way emotional reports were collected during stage 1 required the inclusion of an additional task were the participants had to break the task flow and take a time to type the way the felt during the task. In this stage we aim to perform the sentiment analysis from the texts generated by the task proposed itself (aiming not to break the task flow).

6.2.1.c. *Physiological recording device limitations (data exported time marks and live data)*

Another member of the aDeNu research group has designed and built a physiological data collecting platform based on Arduino. The issues discussed in stage 1 are to be solved as in this stage we have a self-made tool.

6.2.1.d. *Timestamp synchronization*

A tool to generate synchronized timestamps between the computers used to record the data from the experiment has been developed. By mean of this tool, data from a single user collected from two different computers can be synchronized.

6.2.1.e. *Interaction devices usage*

Other open issue from stage 1 was the few interaction data collected due to the design of the task proposed. That issue has been taken into account in this stage and a new essay writing task has been proposed in this stage.

6.2.1.f. *Class attribute format and discretization*

The class attribute processing performed in stage 1 was quite limited because of the format of the labels collected as well as for the labeling methodology followed (with different labelers, some of them providing different formats of labeling). In this stage (also looking for a realistic scalable approach), external labelers are not going to be used. The discretization approach is also something that will be strongly taken into account as the use of different discretization approaches is going to be used as a methodological variable.

6.2.1.g. *Inter-subject approach*

The experimental approach followed in this stage 2 is also based in an inter-subject approach as in stage 1. This time, this issue has been taken into account, aiming to normalize the interaction data to get rid of the data differences from keyboard and mouse due to the variance of interaction skills in the experimental group. This has been done by means of proposing (in section 6.2.2) an interaction baseline, following the approach of the physiological baseline used in the stage 1.

6.2.2. Interaction Data Normalization Approach

In building affective state users' models from keyboard and mouse there is related evidence showing that models which focus on a user's individual interaction patterns tend to be more accurate [122]. To take this into account, detection methods have to characterize the individual features of a person [65]. Modeling an individual person's behavior from keyboard and mouse has several challenges, such as the lack of large interaction data sets in learning settings from which to get an accurate model of the learner [122]. Individuals might have unique keystroke-level reactions to different emotional states [76]. Therefore, although searching for general affect interaction patterns is relatively successful [120], getting personal patterns is more challenging because the accuracy of the methods used is strongly related to the size of the available samples [92].

Bearing in mind the above-mentioned advantages of using individual interaction patterns and following a common practice in experiments using physiological data sources [135,173], we aim to explore the effects on affect detection process and results of an interaction baseline model. This interaction baseline model is based on how the participant interacts with the keyboard and mouse when the user is framed in a setting where it is assumed that they are not affectively involved.

This baseline model establishes a reference model of how the user interacts with the keyboard and mouse, and it has been designed to be obtained from an initial task specially designed for that, called calibration task of from the previous task (see section 6.2.3). From the modeling viewpoint, the purpose is to take advantage of individual user features (in terms of interaction dynamics) and general features [76,120] to identify additional modeling opportunities. In doing so we are taking advantage of modeling each individual´s interaction dynamics [122], while we are addressing the common problem of shortage of individual data [92]. In addition, this approach allows us to consider not just significant affect values but transitions among them, i.e., with respect to the reference temporal point the baseline has been calculated.

6.2.3. Reference baseline for overall normalization

As stated in the previous point, a new approach for interaction data normalization is to be used. Nevertheless, normalization is not only to be applied over interaction data, as it has to be applied over the rest of the data sources (e.g. in stage 1 data normalization was already applied to physiological signals) following the approach introduced in [173]. That is why the inclusion of a calibration task for interaction data sources (a kind of baseline for those data sources, already introduced in previous section) and a baseline for physiological signals (as done in stage 1) has been proposed.

It should be pointed that the inclusion of this baseline throws many variables such as that reference temporal point the baseline is calculated. Although the reference values can be collected from the beginning of the experiment, it would also be interesting evaluating a dynamic approach of the baseline, comparing the interactions during one task with the interactions performed in the previous task. For that reason, we aim to evaluate the data collected following three different approaches regarding the normalization process:

- Raw data with no normalization performed.
- User-normalized interaction data using as reference values the interaction values collected before the first task, during the calibration task in case of interaction data sources and class attribute or baseline task in case of physiological data sources (fixed baseline approach).

- User-normalized interaction data using as reference values the interaction values collected during the previous task to the task that is being processed (dynamic baseline approach).

Regarding the calibration task proposed, also some methodological variables arise when designing how those reference values should be recorded in a task designed for it. In order to generate the baseline model a reference text is usually used to get an individual interaction pattern which compares users' performance over different tasks [76,228]. The choice of text may affect the quality of the initial model in different ways and we need to consider several variables: text length may affect the usability of the approach (long texts are to be avoided as they may distract learners from their regular learning task), the degree of verbosity may affect the richness and variety of referenced features in the model [28].

Nevertheless, as physiological signals are collected both in the initial baseline task, in the final baseline task as well as in the interaction data calibration task, additional reference points for data normalization were also used for the physiological data: i) initial baseline task, ii) final baseline task, iii) a combination of the initial and final baseline tasks, and iv) calibration task.

6.2.4. Data preprocessing

Another point to evaluate in this work is the impact of some preprocessing techniques commonly used for dimensionality reduction and class balancing in datasets. As we have seen in Table 2, it is a common problem the lack of big datasets in many related works. The complexity of the process of data collection (as we have seen in stage 1) hardens the creation of datasets with many instances (requiring for that experiments with many participants). Also, the trend of generating as many features as possible from the data sources proposed in related works, results in datasets with few data instances and many features. As seen in section 4.8, we dealt with the so called "curse of dimensionality", and it is a common problem in data mining scenarios where not many data instances are available. Although in stage 1, it was addressed by filtering highly correlated features, there are other techniques that can be used in data mining scenarios in order to reduce the dimensionality of the dataset (by means of discarding the less "useful" features) [138] and to balance the distribution of the class attribute (in order to generate more robust models) [50]. Although some works have used some techniques with this goal, it is not a common practice the evaluation of the impact their use might have on the results from the generated models. That is why in the current stage, we are to use some techniques in this direction and evaluate the benefits and handicaps of using them.

6.2.5. Task and Emotion Elicitation Method

As our goal is to evaluate this approach in a real-world scenario, a real-world task was chosen. In order to collect interactions with using a keyboard, an essay writing task was designed (similar to those in previous work [28,42]), where keyboard interactions are mandatory and mouse interaction would be needed to edit the text as well as to navigate through the application. An emotional elicitor is also needed to record different affective states from participants. Standardized emotional stimuli employed in other studies such as sounds [40] or images [127] were not used in our research as they are not present in real-world educational scenarios. Instead, as done in stage 1 experimentation, task difficulty and time limits were chosen to elicit different emotions from the participants [134,244]. English as second language in a classroom was chosen as the context for our experiment, as it enables us to manipulate the desired difficulty level of the materials within the context of the scenario itself. Our study attempts to take advantage of both within-subject and between-subject approaches to data collection and analysis. On the data collection side, our naturalistic, between-subject experiment would generate few data instances from each participant, which makes it more difficult to get an accurate set of features from each participant than in within-subject experimental approaches, in which more interaction data are considered [122].

6.2.6. Labeling Approach

As mentioned in section 6.1, one of the main goals of this experiment is getting as close as possible to a real-world scenario. To get there in this experiment, and due to the infrastructure requirements that an external annotator entails, no external affective annotations were used. This way, the only emotional labeling to be used is going to be based on the Self-Assessment Manikin scores self-reported by the participants. Based on that data, different labeling approaches are to be valuated according to two variables:

- User normalization: depending whether the user normalization approach described in section 6.2.2 is used or not.
    - In case the user normalization is used, the labeling will be based on the comparison of the current affective labels and the affective labels used as reference according to (6.1). In this case the labels represent the change or transition from the affective state present at the reference point to the affective state present in the current time.

$$Label_{ij} = \frac{SAM\ score_{ij}}{Reference\ SAM\ score} \qquad (6.1)$$

        - In case the fixed baseline is used, the emotional labels collected in the calibration task will be used as reference values.
        - In case the dynamic baseline is used, the emotional labels collected in the previous task will be used as reference values.
    - In case the user normalization is not used, the raw SAM scale scores given by the participants will be used as emotional labels. In this case the labels represent the current affective state of the participant.
- Discretization method used: as the scores are going to be discretized (following the approach introduced in stage1), two different discretization approaches are going to be used:

  - In case the user normalization is not used (the labeling is based on the raw values given by the participants in the SAM scale, i.e. values from 1 to 9), two different discretization approaches are to be evaluated, depending on the values to be considered as neutral:
    - 1-3: negative; 4-6: neutral; 7-9: positive.
    - 1-4: negative; 5: neutral; 6-9: positive.
  - In case the user normalization is used (the labeling is calculated according to formula (6.1), i.e. values from 0.1 to 9), two different discretization approaches are to be evaluated:
    - <1: negative transition; 1: neutral transition; >1: positive transition.
    - <1: negative transition; ≥1: not negative transition.

6.2.7. Clustering

In this stage, further data processing techniques have been proposed. Due to the high dimensionality, an initial approach evaluating the inclusion of clustering techniques in order to get rid of variables grouping the different data instances into clusters is going to be evaluated. To do that, clustering techniques will be used with the data from the different data sources, generating a different clustering for each data source (providing the clustering algorithm all the variables from that data source).

6.2.8. 2-step classification approach

During the transition stage (described in section 5), a new prediction approach was introduced. This approach was based in 2 different prediction steps: the first one aimed to predict whether there is a non-neutral affective state or not and the second one, performed only on those cases where a non-neutral affective state has been predicted, aims to predict the affective state.

6.2.9. Model Generation Algorithm

As done in stage 1, the algorithm to be used for the model generation is going to be another methodological variable to evaluate in this work. This time the approach is not going to suffer many changes as a set of different data mining algorithms are going to be used to generate the affective models of the learners. The implementations of the algorithms used in this stage (already described in section 4.8) are:

- J48
- Naïve Bayes
- Random Forests
- SMO
- Bagging
- Bayes Net

6.3. Context

As one of this stage's main goals is the evaluation of the approach here proposed in a real educational scenario, an experiment to be hold in a real classroom was set up. The experiment was held in a school in Madrid during April 20th and 21st 2016. The participants were 15-16 year old students. As in the previous stage experiment, the goal was to generate a dataset of affective information during educational tasks. During that experiment, sets of four students could participate at the same time, as 4 set ups were configured in the school's computer laboratory.

From November 7th to 18th 2016, the same experiment was carried out, this time in the aDeNu laboratory in the frame of the Madrid's Science Week, where people of all ages could come and participate in the experiment. Although the task was designed for the participants in the experiment held in April, it was also appropriate for the general public.

6.4. Participants

There were a total of 41 participants in this second stage. 27 participants were recruited for the first experiment (April 2016): 10 male and 17 female, avg. age 15.41. 14 participants were recruited for the second experiment (November 2016): 7 male and 7 female, avg. age 44.35.

6.5. Design

The design of this second stage has been carried out around some of the open methodological issues found in the field (the ones described in section 6.2) with the support provided by one psychologist[15] on related issues. These methodological questions include obtrusiveness, emotional modeling, data preprocessing techniques used and the inclusion of an interaction baseline model. The methodology followed has been developed summarizing the steps reported in related works: i) collect data, ii) provide affective labels for the data, iii) prepare the data and iv) generate predictive models. That approach was already followed in the first stage but some changes have been introduced in this stage. The main changes relate to the data preparation step, where we focused on a different set of methodological issues and explicitly took into account the problems identified in the diverse experiments carried out [192,194,211,212]. These issues, depicted in Figure 33, are: i) the creation of an interaction baseline model, which refers to how the participant interacts with the keyboard and mouse. This allows us to avoid the bias in data derived from including different skill level typing skill level for each user. To create this model, an initial calibration task has been included to calibrate the generation of keyboard and mouse features as well as the affective labels used. In our experiment we evaluate this contribution comparing a user-normalized dataset (based on the comparison of the baseline and the participants' actual usage over different tasks) and a raw dataset

[15] Mar Saneiro

(including only the participants' actual usage over different tasks), ii) the discretization method to transform collected data from a dimensional numerical emotional affective model (i.e. Self-Assessment Manikins) to a dimensional categorical model as described below, iii) the preprocessing and dimensionality reduction techniques commonly used when preparing the data from high dimensionality data sets using different data mining algorithms to generate the model.
While in stage 1 keyboard was only used at the end of the tasks, this study involved three different keyboard-centered essay-writing tasks in which learners were asked to label their own emotions. This change provides more keyboard interactions, thus enabling the creation of a more robust users' model from a larger dataset.

The details of the features included in Figure 33 depict the variables found and steps followed to design the experiment, collect and analyze data.

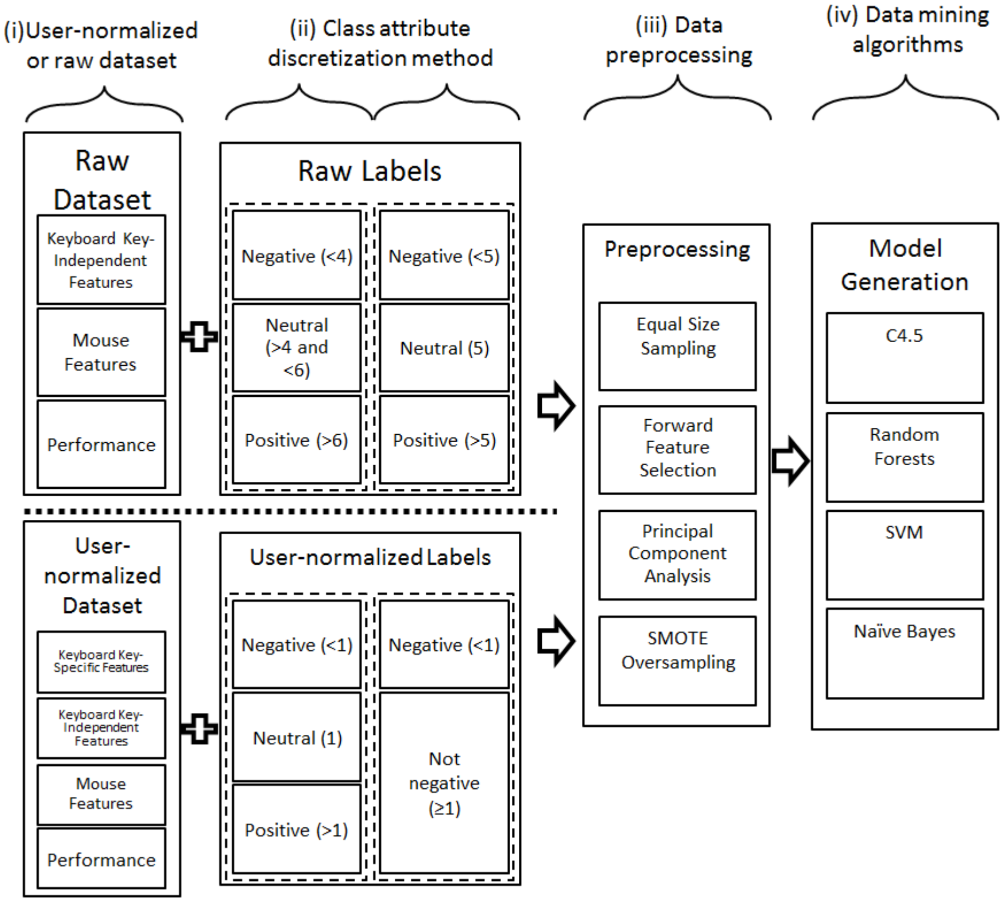


Figure 33. A brief representation of the different methodological aspects evaluated in stage 2 : i) comparing user-normalized dataset (comparing the user's interactions in each task with their interactions from the reference baseline) and raw dataset (including only the participant's interactions in each task). ii) Comparing different approaches to discretize the affective labeling. iii) Different preprocessing techniques used with the data and iv) the data mining algorithms to be used.

6.5.1. Data Sources

Regarding the data sources all the data sources used in stage 1 have been included in this stage. As it was seen in stage 1, the combination of all the data sources provided the best results, so that approach has been followed in this stage. Nevertheless, some aspects of the data sources have been updated:

6.5.1.a. *Keyboard*

Although the keyboard approach is quite similar to the one proposed in stage 1, one of the main problems found in that stage was the lack of use of keyboard due to the nature of the proposed task. In this stage, promoting the use of keyboard has been taken into account in the task design process.

6.5.1.b. *Mouse*

In this stage, the use of mouse during the experiment follows a similar approach to the one proposed in stage 1. The main purpose of the mouse use in the experiment proposed in this stage is navigating through the different tasks. Nevertheless, participants might also use the keyboard in order to select or navigate through the essay they are writing.

6.5.1.c. *Physiological signals*

Regarding the physiological signals, the same signals proposed in the stage 1 have been used. The main change in this aspect is the device used in order to collect those signals. An open hardware-based device was used this time. Within the aDeNu research group, another new research project was initiated, called AICARP[16], which aimed to develop an open hardware based platform designed for sensing the users' physiological state and reacting accordingly with multisensorial feedback using Ambient Intelligence [213,215]. For this research work, only the data recording functionality of AICARP has been used, aiming to provide a low-cost solution capable to provide access to the data recorded in real time. This solution is based on the combination of the e-health platform together with arduino boards, improving some aspects, and includied some adaptations to the experimental limitations introduced by the experiment here described. Nevertheless, the inclusion of some adaptations performed in the AICARP platform were proposed by this work in order to make it less intrusive, as it was a requirement in order to be used in the experiment described in this stage. That is the case of the inclusion of a photoplethysmography sensor to be used in the ear lobe instead of the initial approach of putting the photoplethysmography sensor in the finger (as using that sensor in the finger may interfere on the normal typing performance of the participants). Additionally, thanks to the data analysis performed in this Thesis, some design issues were found and reported for their solving.

The platform consists on a central module, which is connected to the different sensors and the computer where the data is going to be stored. A tool has been developed in Matlab in order to control the platform and manage the data collection,

[16] Although this platform has been used in the experiments of stage 2 and it is described here, the development of this platform (both the hardware and the software described in section 6.5.1.c) has been carried out by other member of the aDeNu group (Raúl Uría Rivas supervised by Jesus Boticario and Olga Santos) so this hardware-based platform used in this stage is not a contribution resultant from this PhD thesis (but some changes were suggested by this work and carried out by Raúl as well as some errors on the data quality of the platform were fixed by Raúl thanks to the reports provided by this work).

recording and visualization. Regarding the sensors, the following physiological sensors were used in stage 2:

- Heart rate sensor
- Breath sensor
- Skin temperature sensor
- GSR sensor

6.5.1.d. *Facial expressions*

As done in stage 1, the facial expressions were recorded via webcam, although their use is left for future works. The Kinect sensor was discarded for this stage as it was discontinued by its manufacturer.

6.5.1.e. *Sentiment analysis*

Another benefit from the new task proposed in this experiment is the capability of performing sentiment analysis in a less intrusive way. As mentioned in stage 1, in order to perform sentiment analysis, a task where the participants were asked to type down their emotions was included. That task can be considered intrusive as introducing a task in order to exclusively detect the affective state of a learner might disrupt her learning flow [REF]. In this stage we aimed to improve the sentiment analysis approach in some points:

- In order to get closer to a real-world scenario, no additional tasks were included to perform sentiment analysis. The sentiment analysis in this stage has to be done from the text collected in a real task.
- The approach followed in this stage aims to generate its own sentiment analysis model for the tasks proposed, in contrast to the approach followed in stage 1 where a lexicon was used. This point will be further discussed in section 6.7.4.

6.5.2. Labeling

Labeling is one of the methodological variables that have been simplified in contrast with the approach followed in stage 1. Although an evaluation of several labeling approaches (and sources) was carried in stage 1, in this stage 2, one of the main goals is to transfer the stage 1 experimentation to a real-world context. That goal makes it hard to think of a real-world approach where there are external labelers capable to label interactions from groups of e-learners in real time (although there are methodologies for live labeling groups of students [REF BROMP], that approach is not applicable in distance e-learning scenarios).

6.5.3. Tasks

The design process was carried out meticulously in order to find a perfect balance between the goal of the experiment and the realistic fitness of the proposed task in a real-world educational scenario. In order to find that balance, the task chosen was essay writing in the frame of the subject English as a Second Language. Essay writing is a task that is commonly carried out in Second Language Acquisition, so it can be evaluated perfectly in a real classroom. For this experiment, the materials being used by the classroom were used in order to adapt the task proposed [37].

The experiment consisted of 3 different tasks, with each task consisted in writing an essay with a series of vocabulary terms proposed to be included in the essay. The choice of the vocabulary terms proposed was used as a way to elicit emotions, using terms from lessons the participants had already seen in the first two tasks and material from an still unseen lesson in the last task. By means of this, we aim to induce frustration and stress. As done in stage 1, time limits were also used as emotional elicitators, having a time limit of 210 seconds for each essay to write. In this stage, in comparison with stage 1, another variable was introduced to try to elicitate stress to the participants: while participants could read the proposed vocabulary terms to use in the first essay. In the second and third task they had to memorize the terms to use in the essay. In those two last tasks, participants were shown the proposed terms to be used in the essay during 30 seconds. Once the time was over, the terms disappeared from the screen and then participants were allowed to type.

6.5.4. Infrastructure

The stand for each participant was redesigned, using for the experiments in this stage only two computers per stand. This makes our stands to be more portable as well as allows us to use more stands (as less computers are needed). This is in part, due to removing the use of the Kinect device. It also should be noticed that new computers were used. Another important change was the use of the AICARP platform, removing the J&J device used in stage 1. This change on the device for collecting physiological signals added some improvements to our design in the following 2 points: i) it was developed by people in the research group, so any required customization could be implemented (e.g. changing the initial HR sensor, to be placed in the finger, for other to be placed in the ear lobe, so participants could type better) and ii) the webcam video was also recorded by the AICARP tool, which means one less program running on the computers.

Computer	Used by	Running software	Devices attached
Participant's computer	Participant	• MOKEETO tool: to perform the tasks proposed (to be presented in section 6.5.6.c). • Keylogger/mouse tracker app (implemented): to record the participant's interactions with keyboard and mouse. • Remote desktop program (VNC): to allow the participant's tutor be aware of the advances of the participant during the experiment, allowing him or her to take a timestamp every time the participant ends a task. • Synchronization application: to keep a track of the time differences between the two computers used in the experiment.	none
Tutor's computer	Tutor	• AICARP tool: to record the participant's physiological signals as well as the webcam recording. • Synchronization application: to keep a track of the time differences between the two computers used in the experiment. • Remote desktop program (VNC): to allow the participant's tutor be aware of the advances of the participant during the experiment, allowing him or her to take a timestamp every time the participant ends a task.	• Webcam (Logitech C310 or Quickcam Pro 9000) • Physiological sensors (AICARP)

Table 26. Configuration of the computers used in each stand in stage 2

6.5.5. Materials

The materials prepared to perform this experiment are listed next.

- ***Information consent:***
  As done in stage 1, an information consent was required in order to record the data from the participant. This time, to get the information consent from the students in the school, it was sent to their parents (as the students were not legally allowed to sign it) prior to the experiment, so only those students whose parents signed it were allowed to take part in the experiment.
- ***Task materials:***
  The choice of the proposed words was made from the learning materials followed by the class that participated in the experiment. The vocabulary was extracted from the book Activate B2. The proposed words can be found in Appendix II (in section 13.2.9)
- ***Demographic information:***
  Some questionnaires to get some demographic information.

6.5.6. Implementation

The technological infrastructure prepared for the experiment included the keylogger and mouse tracker application already used in stage 1. Two new tools were developed for this experiment: a data synchronization tool and a tool for writing essays with different possible configurations (proposed words, time limit, etc.).

6.5.6.a. *Keylogger and mouse tracker application*

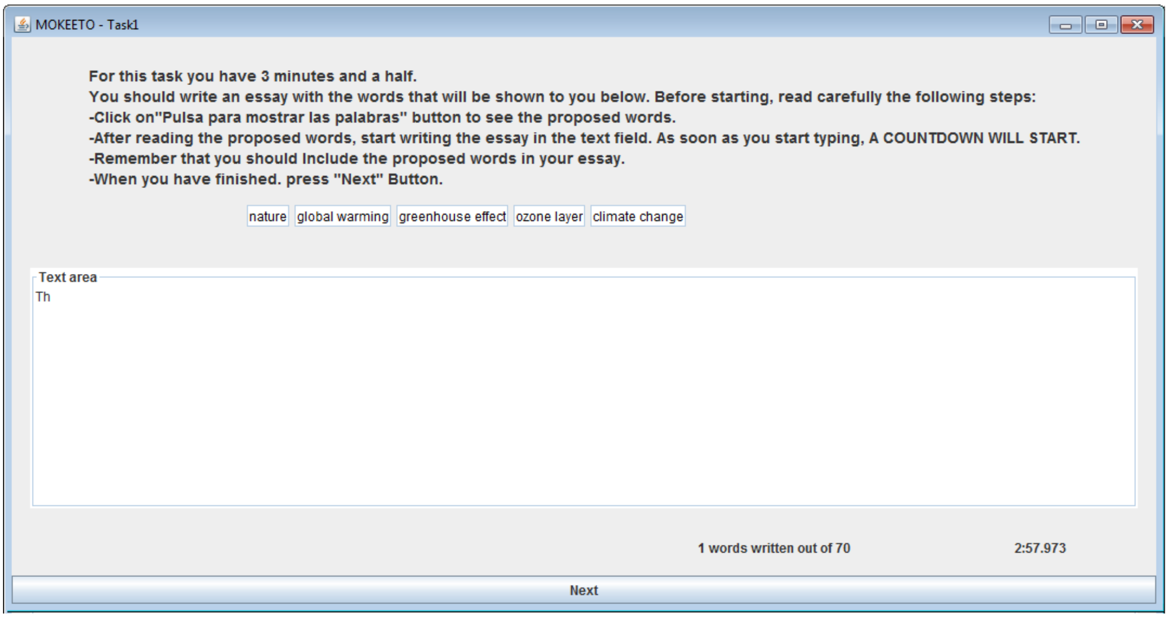
The tool used in stage 1 to record mouse movements and keyboard interactions was also used for this experiment. Two new characteristics were added for this version of the logger: i) the tool now can record the mouse scroll interactions and ii) it also records the program being shown in the foreground of the OS (the tool the participant is interacting with). Nevertheless, due to the nature of the task proposed, these two features were not used (as scroll interactions were not necessary and the participant did not have to switch between different programs).

6.5.6.b. *Synchronization application*

Due to the new tool being used for physiological signals recording worked as a standalone tool in a different computer, a program to take timestamps was developed in order to save a log with the time of the different computers involved in the data collection (the participant’s one and the experimenter’s one). By means of running the program in both computers (the program is invisible and runs in the background) and pressing a predefined key, one computer sends its timestamp to the other, which keeps that time stamp together with its own timestamp. By mean of this, physiological signals log will be able to be synchronized with the rest of data sources’ logs.

6.5.6.c. *Essay writing tool*

An essay writing tool called MOKEETO (MOuse and KEyboard logging Essay writing TOol) was implemented, in order to log all user interactions with their corresponding timestamp. The tool consists of a sequence of panels (each one corresponding to a different task) with three main sections: the task instructions shown on the top of the screen, a text input form in the center of the screen and a set of indicators (written words counter and a timer showing the time left in real time) at the bottom of the screen (see Figure 34). Additionally, each task can be configured to: i) show in the task instructions section (without allowing copy/paste) a set of words to be included in the essay, ii) hide that set of words after a given time, disabling the text input form while the words are being shown (so the participant has to remember the words before starting to type the essay, which may increase the difficulty of the task), and iii) prevent the participant from skipping the current task until a given set of circumstances are given. In our experiment the participant couldn’t skip the task until more than 70 words were typed or a 210 seconds time limit was over).

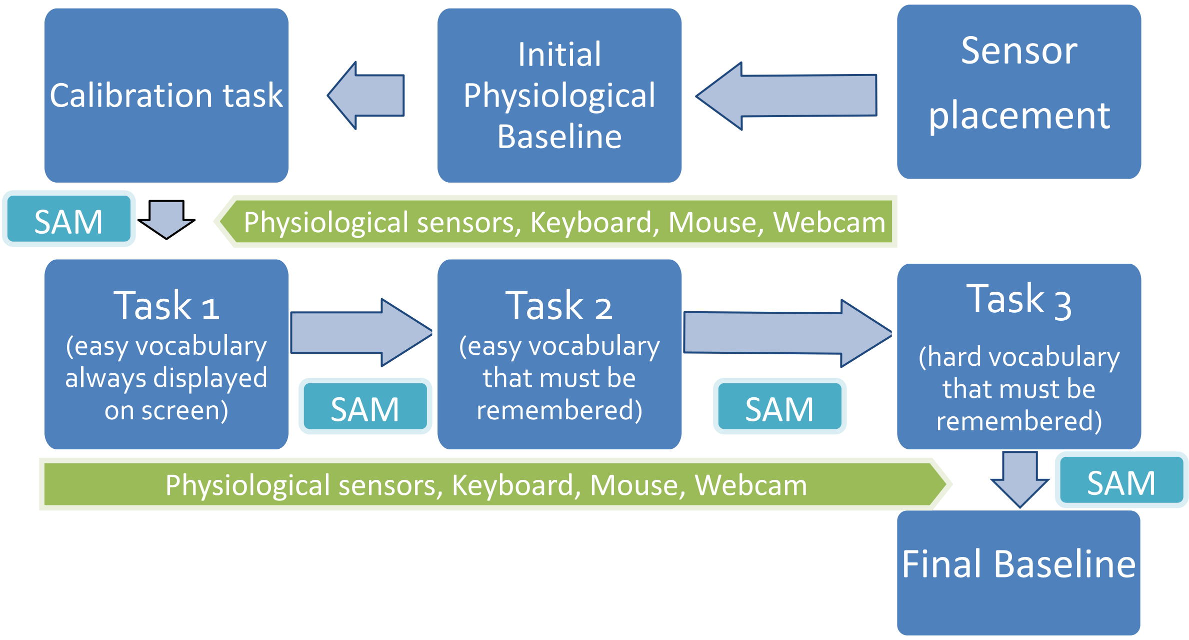


For this task you have 3 minutes and a half.
You should write an essay with the words that will be shown to you below. Before starting, read carefully the following steps:
-Click on"Pulsa para mostrar las palabras" button to see the proposed words.
-After reading the proposed words, start writing the essay in the text field. As soon as you start typing, A COUNTDOWN WILL START.
-Remember that you should Include the proposed words in your essay.
-When you have finished. press "Next" Button.

Figure 34. MOKEETO: Essay tool used in the experiment. Instructions are given at the top of the screen, proposed words are shown below task instructions and the text area is shown in the middle of the screen. Word counter and time remaining are shown at the bottom of the screen.

6.5.7. Procedure

The experiment was structured in three different tasks (or essays), with an incremental difficulty. Previously participants were asked to go through an interaction baseline.



Sensor placement
Initial Physiological Baseline
Calibration task
SAM
Physiological sensors, Keyboard, Mouse, Webcam
Task 1 (easy vocabulary always displayed on screen)
SAM
Task 2 (easy vocabulary that must be remembered)
SAM
Task 3 (hard vocabulary that must be remembered)
Physiological sensors, Keyboard, Mouse, Webcam
SAM
Final Baseline

Figure 35. Stage 2 experimental structure, including tasks and data to be collected.

We can see how, in order to adapt the AMO-ML methodology to a more realistic context, the experimental design has been simplified in contrast to other previous experiments (shown in sections 4.5.7 and 5.2.5). In order to carry out that simplification, some stages have been removed (e.g. self-emotional report) in order to reduce possible disturbances in the normal development of the proposed task.

6.5.7.a. *Part 1: welcome and preparation*

The first part of the experiment consisted of a series of steps designed to set up the recording devices to be used:

- Sensor placement and recording. All the recording devices were set (if not running yet) to start recording:
    - Heart rate sensor: Participants were attached a heart rate sensor in the earlobe as a no ear hole earring.
    - Respiratory sensor: A belt was tied around the participant’s chest in order to registry the volume of air consumed.
    - Skin conductance sensor: Two velcro straps to be placed on the index and ring fingers of the non-dominant hand of the participant.
    - Temperature sensor: A sounding was placed in contact to the participant’s wrist attached by a wristband.
    - Screen recording: The program Camtasia was configured in the tutor computer previously to the arrival of the participant and started recording prior to the start of the experiment.
    - Mouse tracker and key logger: The program developed to record mouse and keyboard interactions was launched before the participant entered the room.
    - Synchronization tool: The program developed to collect timestamps of the tutor computer and the participant computer in order to synchronize the data collected in both computers was launched in both computers before the participant entered the room.
    - Remote desktop: The program VNC server was set up to allow the participant’s tutor view the participant’s screen all along the experiment.
    - Physiological signals and webcam (AICARP tool): the tool developed to record the physiological signals and webcam had to be set up to start recording.
- While the sensors were placed, some demographical questions about gender, age, computer skills, information that may affect some sensors measures (sports, smoking and medicines) were asked.
- Initial base line. Participants were asked to relax for 2 minutes in order to get the values of their physiological signals while relaxed.

6.5.7.b. *Part 2: task*

After the physiological baseline was recorded, participants started to interact with MOKEETO. Because the purpose was to detect participants’ affective changes, bearing in mind the relationship between affect and cognition demands discussed in previous related research, several factors are used in the following tasks to increase the difficulty over time. These include: i) time limit (as used in [137]), ii) proposed words difficulty (proposing uncommon words in the last task and common words in the first tasks) and

iii) forcing the participant to remember (instead of viewing all over the task) the required proposed words. Each task had an introductory small text with the instructions of each task so participants knew what they had to do before each task. With these factors in mind, the following tasks were proposed:

- Task 0 (calibration task):

  The initial task (from now on, calibration task) was designed to create a base model of how the participant interacts with the keyboard and mouse to be used as a reference (or baseline) when comparing the keyboard and mouse interactions of the following tasks. The use of a baseline has been traditionally used in experiments using physiological data sources. Our approach has adapted this to the data sources proposed (keyboard and mouse) and its usefulness is one of the methodological questions we aim to evaluate (see hypothesis **H2** in the introduction section). This baseline model, aims to be obtained in a scenario where the user is not affectively involved, so participants were proposed to copy a short excerpt from Alice in Wonderland (as done in related works [76,228]).

- Task 1:

  The first task was designed to be easy. Participants were asked to write an essay with 5 proposed words. This task had the following conditions:

  - The proposed words were chosen from a lesson the students had already seen in their class. The selected words were common words.
  - The proposed words are shown all over the task, so participant does not have to memorize anything.
  - Participants had a time limit of 210 seconds

- Task 2:

  The second task was similar to the first one, although some variables were changed in order to make it a little bit harder. Participants were asked to write an essay with 5 proposed words. This task had the following conditions:

  - The proposed words were chosen from a lesson participants were studying at the time of the experiment.
  - The proposed words were shown at the beginning of the task. The words were shown only during 30 seconds (so participants had to memorize them). Once the words disappeared from the screen, participants were allowed to start typing.
  - Participants had a time limit of 210 seconds

- Task 3:

  The third task was designed to be the hardest one:

  - The proposed words were highly uncommon and taken from a lesson the participants still haven't studied.
  - The proposed words were shown at the beginning of the task. The words were shown only during 30 seconds (so participants had to memorize them). Once the words disappeared from the screen, participants were allowed to start typing.
  - Participants had a time limit of 210 seconds

After each task (including the calibration one), the participant was asked to express her affective state. Self-report is one of the most common approaches followed to determine affective states as seen in most works analyzed in [120]). This self-report was given by means of the Self-Assessment Manikin (SAM) scale [39], providing a score, from 1 to 9, for both valence and dimensions of their current state at the end of each task. The SAM scale was presented with a textual explanation of each one of the affective dimensions to label. Valence and arousal values will be used to generate the affective attributes to be predicted by the system.

6.5.7.c. *Part 3: post-experiment*

The third and last part ended the experiment, and collected the participants' baseline at the end (i.e., participants are asked again to relax for 2 minutes), removed the sensors from them and asked them some feedback with the following questionnaires:

- Positive and Negative Affect Schedule (PANAS): a 10 item questionnaire to measure the primary dimensions of the mood [203,243].
- Satisfaction questionnaire: to allow participants report if they liked the experiment.

Once the experience ended, participants were debriefed about the experiment and allowed to ask as many questions as they wanted. They were also shown how the information was recorded and why.

6.6. Data recorded

As done in stage 1, the data had to be processed after the experiment. This time, the data for each participant was named after the experimental stand and the number of participant in that set:

puestoX_userY

Being:

- XX: the number of the stand the participant was seated on ({1,2,3 or 4} in case of the experiment in the classroom or 5 in case of the experiment in the science week)
- YY: the number of the user in that stand

The data files generated for each participant included information from the different devices used in the experiment. Details are provided next.

6.6.1. Webcam video

The result of the webcam recording was a .avi file containing only the video (with no audio track) with a 320x240 resolution and 10 frames per second. The files generated

As aforementioned, this information was not used in the research work reported in this Thesis, but is included here for completeness and to support future work.

Figure 36. A frame of a video recorded during the stage 2.

6.6.2. Webcam Audio

The audio of each session is also recorded by the webcam microphone and saved in a file. The file contains the audio recorded in mono, 8000Hz at 16 bit.

6.6.3. Keyboard interactions

The file generated by the keylogger/mouse tracker app containing the keyboard interactions is a csv file with the following information:

- Time of the event: hour, minute, second and millisecond
- Type of event: ‘p’ for press or ‘r’ for release
- ASCII code of the key
- Representation of the key
- The name of the active window
- The name of the process for the active window

Here is shown a log extract to see the fields generated:

```
10:54:47:578;p;13;RETURN;logs;Explorer.EXE
10:54:47:671;r;13;RETURN;logs;Explorer.EXE
11:30:31:375;p;20;CAPITAL;Composition Tool - copiaTexto;javaw.exe
11:30:31:515;r;20;CAPITAL;Composition Tool - copiaTexto;javaw.exe
11:30:32:078;p;65;A;Composition Tool - copiaTexto;javaw.exe
11:30:32:234;r;65;A;Composition Tool - copiaTexto;javaw.exe
11:30:32:531;p;20;CAPITAL;Composition Tool - copiaTexto;javaw.exe
11:30:32:640;r;20;CAPITAL;Composition Tool - copiaTexto;javaw.exe
11:30:33:062;p;76;L;Composition Tool - copiaTexto;javaw.exe
```

As the interactions are exported as plain text, the files generated are not extremely huge, depending on the file size on the number of interactions performed during the session. A file with 6055 events registered (in a 68 minute session) takes only 343 kb. This time, due to the inclusion of the active window name and its process, the file takes more space, but still being a small file.

6.6.4. Mouse interactions

The keylogger/mouse tracker app also generated another csv file with the mouse interactions. In this case, the file contained the following information:

- Time of the event: hour, minute, second and millisecond
- Type of event: ‘mv’ for movement, ‘prl’ for left button pressing, ‘rll’ for left button releasing, ‘prr’ for right button pressing and ‘rlr’ for right button releasing, ’scru’ for scroll moving up and ‘scrd’ for scroll moving down.
- X coordinate: coordinate X in pixels starting from the left part of the screen where the event has been registered
- Y coordinate: coordinate Y in pixels starting from the top part of the screen where the event has been registered
- The name of the active window
- The name of the process for the active window

The event log generated looks like this:

```
15:16:14:359;mv;671;763;Composition Tool - copiaTexto;javaw.exe
15:16:14:671;mv;671;762;Composition Tool - copiaTexto;javaw.exe
15:16:14:781;prL;671;762;Composition Tool - copiaTexto;javaw.exe
15:16:14:906;rlL;671;762;Composition Tool - copiaTexto;javaw.exe
15:16:21:203;mv;671;761;Composition Tool - Sam-Valencia-
Tarea0;javaw.exe
```

These files usually need less space in hard disk (depending this on the interactions performed during the session). A file containing 19245 interactions (in a 70 minute session) takes just 1227 kb of disk space.

6.6.5. Physiological signals

The AICARP tool generated a csv file with all the values recorded from the physiological signals. The platform has been developed to take measures with a frequency of 10Hz. Nevertheless, the registries are not exported exactly every 100 ms. This was a problem as in some cases the time between samples in some cases keep growing in some users in the experiment held in April 2016 (this issue description and management has been described in section 6.7.2). The following columns are exported in the csv file:

- Hour: The hour of the system at the time the data was recorded.
- Minutes: The minutes of the system at the time the data was recorded.
- Seconds: The seconds of the system at the time the data was recorded.

- Milliseconds: The milliseconds of the system at the time the data was recorded.
- Index: The number of registry recorded since the beginning of the recording
- State ID: ID to identify if the current registry is in a physiological baseline
- Breath: The data recorded from the breath sensor.
- Soften breath: The data recorded from the breath sensor, softened.
- BCPM: Number of breaths per minute calculated from the data recorded by the breath sensor.
- Conductance: The data recorded from the skin conductance sensor.
- Soften conductance: The data recorded from the skin conductance sensor, softened.
- Temperature: The data recorded from the skin temperature sensor.
- Soften temperature: The data recorded from the skin temperature sensor, softened.
- BPM: Heart rate.

Here is an excerpt from one of the physiological logs recorded:

```
15;59;26;448;102;1;541.000;539.000;40.000;0.000;0.000;0.500;0.510;32.160;32.250;69
15;59;26;546;103;1;541.000;538.000;40.000;0.000;0.000;0.500;0.510;32.310;32.210;69
15;59;26;693;104;1;538.000;538.000;40.000;0.000;0.000;0.520;0.510;31.960;32.210;69
15;59;26;807;105;1;538.000;538.000;40.000;0.000;0.000;0.550;0.500;32.510;32.220;69
15;59;26;926;106;1;542.000;537.000;40.000;0.000;0.000;0.510;0.500;32.040;32.220;69
15;59;27;63;107;1;540.000;538.000;40.000;0.000;0.000;0.490;0.500;32.120;32.210;69
15;59;27;180;108;1;537.000;538.000;40.000;0.000;0.000;0.490;0.490;32.430;32.260;70
15;59;27;280;109;1;540.000;537.000;40.000;0.000;0.000;0.490;0.480;32.190;32.220;70
15;59;27;417;110;1;539.000;536.000;40.000;0.000;0.000;0.470;0.480;32.230;32.250;70
```

6.6.6. Screen recording

The screen of the observer’s computer was recorded using the screen recording software Camtasia with a resolution of 800x600 pixels. As we can see in Figure 37, that recording contains the image from the webcam, the physiological signals and the participants’ desktop.

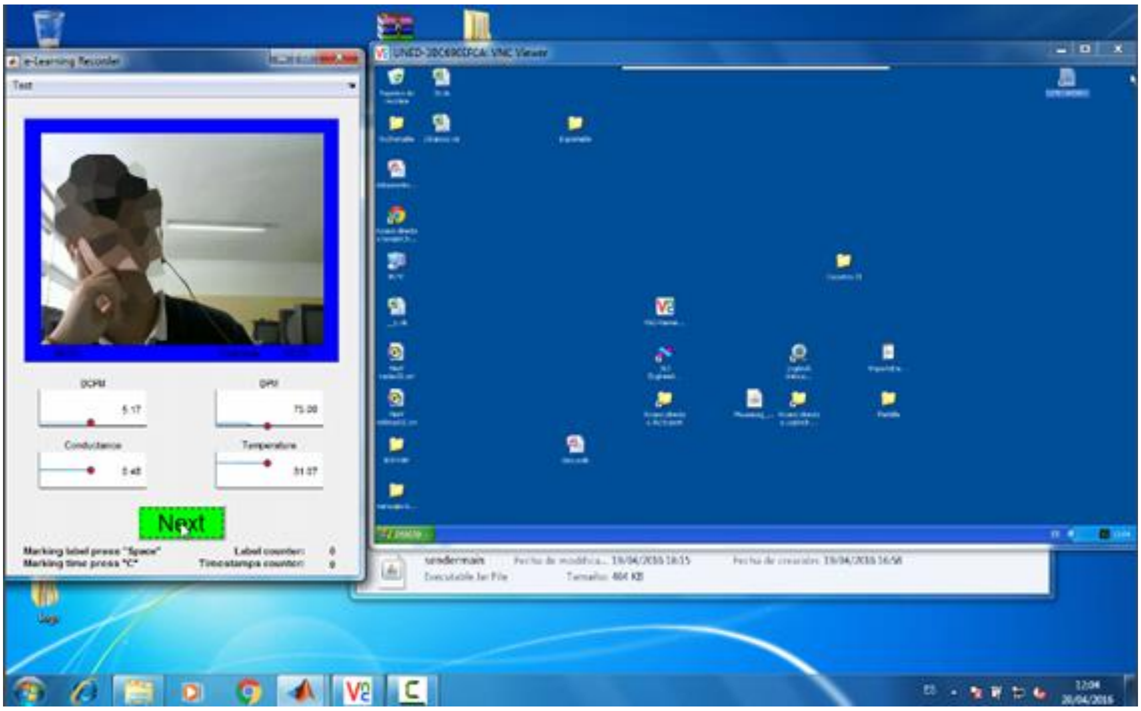


Figure 37. Screenshot of a recording of the observer's computer screen in stage 2. The webcam image can be seen at the left part of the screen as well as the physiological signals recorded live. On the right part of the screen, the participant's desktop is being shown.

6.6.7. Timestamps

As mentioned in section 6.5.6.b, a tool for taking timestamps in order to synchronize the data recorded from both computers was developed. Every time the observer pressed a key, the time from the observer’s computer was sent to the participant’s computer and recorded with the time of that reception by participant’s computer.

```
2016/04/21 14:47:57:124;TIMESTAMP;ServerTime14:47:59:397
2016/04/21 15:04:50:124;TIMESTAMP;ServerTime15:04:52:257
2016/04/21 15:13:21:468;TIMESTAMP;ServerTime15:13:23:637
```

6.7. Data preparation

Following the steps depicted in section 1.5, the data collected has to be preprocessed before generating the data mining models. In this stage, the data preprocessing process has a strong importance, as some of the methodological variables evaluated are related with the way the features are prepared.

Also, as in stage 1, we do not consider the information from the webcam video, so its preparation is not reported here.

6.7.1. Interaction devices models

To model keyboard interactions, key events were recorded (key press and key release time) and processed, generating two different sets of features, with key-specific features and key-independent features. To model mouse interactions (specifically mouse movements), click and scroll events were recorded. Note that although the latter was not used as the graphical user interface, the use of scroll was not actually required, generating only one set of features. To model performance, a particular single set of features was generated. For the mouse, keyboard and affective labeling datasets, two versions were created: one using information from the calibration task or from the previous task baseline (i.e., user-normalized dataset) and another without using that information (i.e., raw dataset), as it is illustrated in Figure 38. The purpose here is to evaluate to what extent using a baseline to normalize regular tasks improves data prediction models.

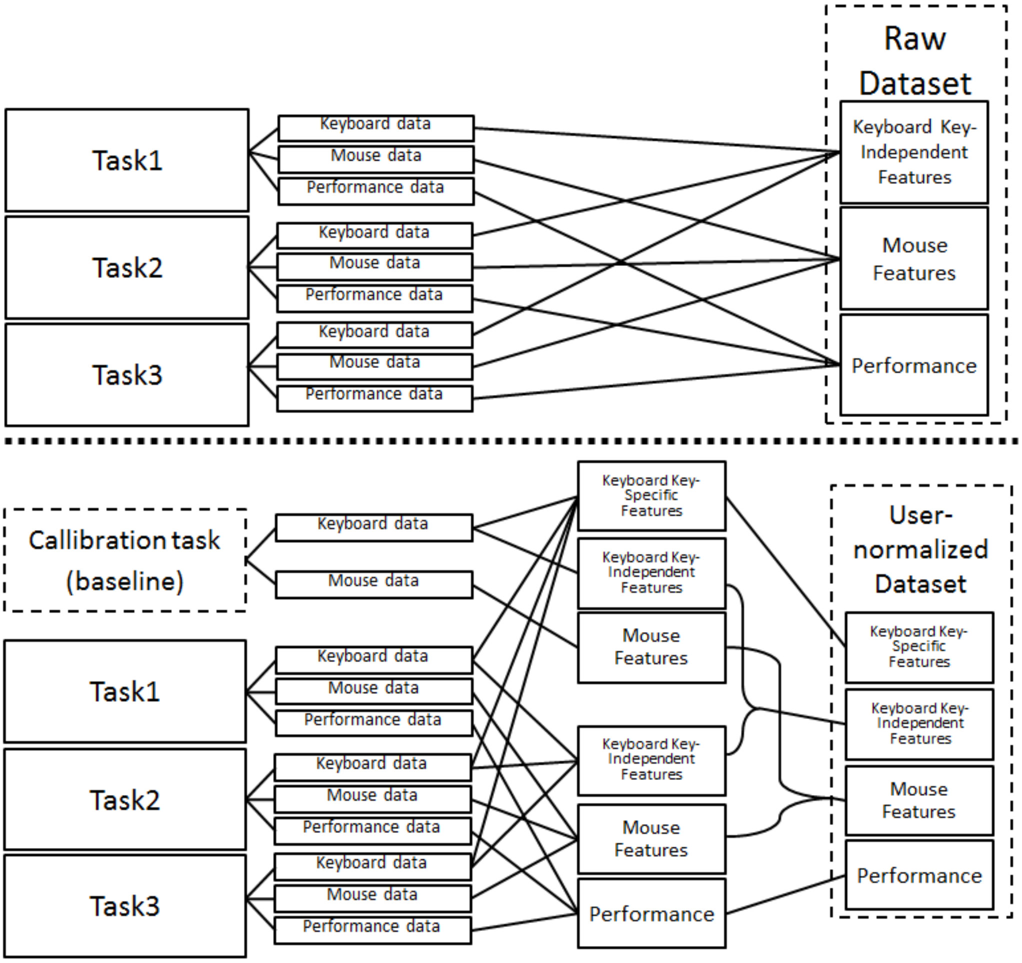


Figure 38. Different approaches in the feature generation process in order to generate the different datasets : the user-normalized (based on the data collected in the calibration task or the previous task) and the raw one. User-normalized Dataset (bottom right) is generated by comparing data collected from the calibration task or previous task (dashed line rectangle on the left) and the regular tasks (bottom left). Raw Dataset (top right) is generated using only data collected from the regular tasks (top left).

The process followed to generate the user-normalized dataset feature values is as follows (values and process are shown in Figure 39). Initially, the raw data (Raw Dataset) is obtained, where each participant has a particular skill level using the keyboard and mouse that is represented by a specific value. In the Data Normalization stage, the values collected in the calibration task or the previous task are used as a reference value for each user. Each value recorded for a participant in every task is divided by the value of that specific participant in the baseline model (i.e., in the calibration task or the previous task). The result of this is used in the final step to build the User-normalized Dataset, where the goal is to get rid of the possible differences in the values among participants due to their respective keyboard/mouse interaction skills. In other words, in normalized dataset, values represent the proportion a feature has changed compared to the baseline value (for that same feature for that given participant). The normalized values were calculated using the following formula:

$$Normalized\ value = \frac{Raw\ value\ calculated\ in\ task}{Raw\ value\ calculated\ in\ user\ baseline} \qquad (6.2)$$

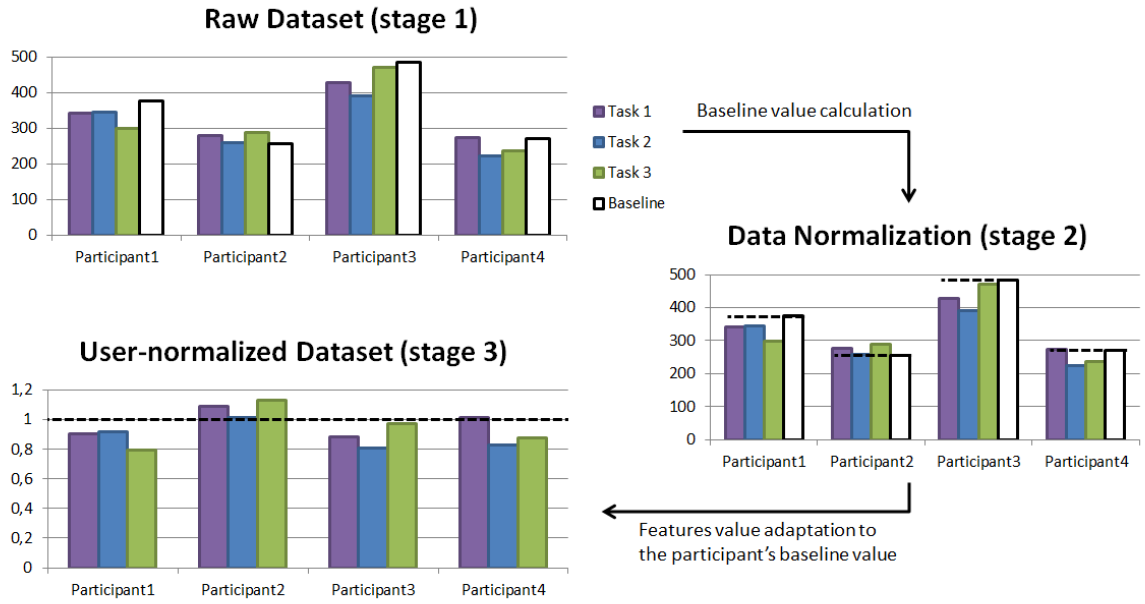


Figure 39. Interaction data normalization process followed. Each participant data is normalized according to the reference value (white column).

From the interaction devices, three different sets of features are to be generated: two generated from the keyboard interactions (one fine grained model generated from for every combination of keys performed, called keyboard key-specific feature model, and a second model taking into account all the keyboard interactions regardless the keys pressed, called keyboard key-independent model) and one generated from mouse interactions:

6.7.1.a. *Keyboard Key-Specific Feature Model*

This dataset aims to model the differences reflected when typing the same keys in two different times (i.e., in different tasks). To be more precise, different ways of typing pairs of consecutive keystrokes (digraph) in a given task compared to the way that same pair of consecutive keystrokes was typed in the initial calibration task or the previous task (This makes this model dependent of a normalization process). To generate that model, first step was to group all the combinations of two and three consecutive keystroke events in every task (including the calibration task). Once all the combinations were created and in order to generate a precise model of the user's typing, only those combinations that were stroked during the same task over a given number of times were kept. In contrast with [42], where no minimum number of instances per n-graph is set, we set up a minimum digraph appearance threshold to 3. Figure 40 illustrates the impact of this threshold in the numbers of digraphs generated for each task and number of digraphs "used" in common with the calibration task or the previous task. As the threshold value increases, the model generated relies on more observed instances of the same digraph, which makes it more solid, but the number of typed digraphs which value were over that threshold decreases. As we can see in Figure 40, in case the minimum number of appearances was set to n=4, the number of digraphs used to generate the user model in each task could go close to 0.

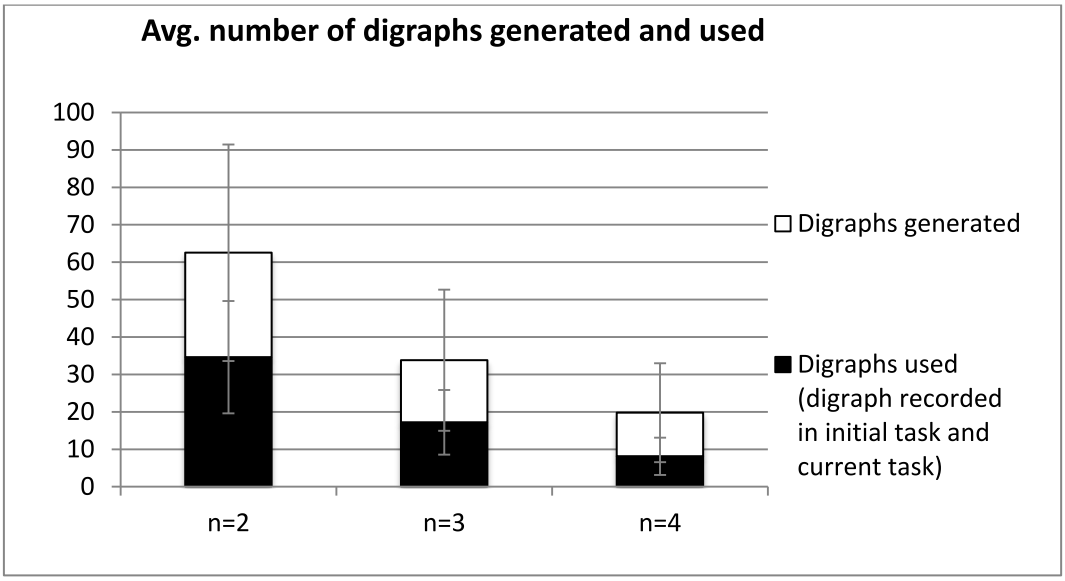


Figure 40. Avg. number of digraphs generated for each task depending on the number of instances (n) of that digraph in the text. Also, number of digraphs (of the same type as those recorded in the calibration task) used for each task.

For these combinations, all the digraph and trigraph features present in [76] were calculated for each task (included the calibration one) following the same coding presented there:

- 2G_1D2D: The duration between 1st and 2nd keypress start times of the digraphs.
- 2G_1Dur: The duration of the 1st keypress of the digraphs.
- 2G_1KeyLat: Duration between 1st keypress end time and next keypress start time of the digraphs.
- 2G_2Dur: The duration of the 2nd keypress of the digraphs.
- 2G_Dur: The duration of the digraphs from 1st keypress start time to last keypress end time.
- 2G_NumEvents: The number of keypress starts that were part of the graph (as possible overlapping may occur).
- 3G_1D2D: The duration between 1st and 2nd down keys of the trigraphs.
- 3G_1Dur: The duration of the 1st key of the trigraphs.
- 3G_1KeyLat: Duration between 1st key up and next key down of trigraphs.
- 3G_2D3D: The duration between 2nd and 3rd down keys of the trigraphs.
- 3G_2Dur: The duration of the 2nd key of the trigraphs.
- 3G_2KeyLat: Duration between 2nd key up and next key down of trigraphs.
- 3G_3Dur: The duration of the third key of the trigraphs.
- 3G_Dur: The duration of the trigraphs from 1st key down to last key up.
- 3G_NumEvents: The number of key events that were part of the graph.

Once we have all the features for each task, we calculate the proportion for each feature compared to the same feature and same digraph in the calibration task or the previous task, following formula (6.2), using that proportion for each feature in the dataset.

6.7.1.b. *Keyboard Key-Independent Model*

The second dataset generated was designed for a not so fine grained keyboard use modeling, based on key-independent n-graphs and task related keys features. The features presented in the previous dataset from [76] were also generated in this dataset. The main difference with the previous dataset is that in this dataset, all the n-graph features are aggregated, while in the previous dataset, only a subset of the n-graphs recorded were used (only those that appeared over a given threshold in the given task and the calibration task).

Also some other features not included in [76] were generated in this dataset:

- Overlapping press events: number of press events occurring while another key was already pressed.
- Uppercase press: number of press events occurring while shift key was already pressed.
- Pauses: time between a key release and a key press events.
- Time between two consecutive Press events: time between two different consecutive key press events.
- Time between Press and Release events: time between the key press and key release events of the same key. This indicator was calculated taking into account different sets of keys (backspace key, backspace and delete keys, delete key, alphabetical characters and space bar).
- Word separation: time between the release event of a character key and the press event of another character key separated with a keystroke of the space bar.
- Typing proportion time: the proportion of the whole task time where there has been a key being pressed.

For this dataset, two different versions were generated in order to compare if the use of the data of the calibration task or the previous task may affect the predictive outcomes, thus generating differential values between the features in each task and user compared to those in the calibration task or the previous task. The two datasets generated were: i) the “raw” version of the dataset containing values for each feature calculated in each task and ii) the “user-normalized” version of the dataset containing, for each user, the comparison between the feature values in each task and the corresponding feature value in the calibration task or the previous task following formula (6.2).

6.7.1.c. *Mouse interactions*

The third dataset was generated from the mouse interaction logs (containing mouse movements, clicks and scroll movements). First step is using an aggregation method to group the events from the raw data recorded to generate the model features. Traditionally, in mouse interaction modeling, events are grouped into mouse movements, but there is not a standard definition of mouse movement: while in [171] the events were grouped into movements delimited by their distance, in [126] is the time of inactivity that is used to separate mouse movements. We built our model following the second approach and all the mouse cursor coordinates recorded during the experiment were grouped in mouse movements, defining a mouse movement a series of coordinates that vary along time with a time difference between each one below a given time threshold (in this case it was set to 500 milliseconds as it was done in [126]). If a change in the cursor position is produced after a time period over that threshold, that position will be considered the starting position of a new mouse movement. After that, the following features (proposed in [248] and followed in our previous work [199]) were calculated:

- Movement accumulated angle variation: the angle variation described by the cursor for every pair of consecutive cursor locations compared to the angle described by the previous pair of cursor locations.
- Average movement acceleration: the average acceleration in each movement.
- Movement acceleration standard deviation: the standard deviation of the acceleration in each movement.
- Average movement speed: the average speed of the mouse movement.
- Movement speed standard deviation: the standard deviation of the mouse speed in a mouse movement.
- Distance covered: the distance covered by the cursor in a mouse movement.
- Euclidean distance: Euclidean distance between the coordinates where a mouse movement begins and the coordinates where the mouse movement ends.
- Difference between “distance covered” by the mouse cursor in a movement and “Euclidean distance” between the starting and the end point of the movement.
- Click covered distance: the distance covered by the cursor while a mouse button was clicked.
- Click Euclidean distance: Euclidean distance between the coordinates where a mouse button was pressed and the coordinates where the mouse button was released.
- Click time: time between a mouse button press event and the consecutive mouse button release event.
- Difference between “click covered distance” and “click Euclidean distance”.
- Pause length: the length of the pauses between mouse movements (>1s).

As in the keyboard key-independent dataset, two different versions, user-normalized and raw, were generated from this dataset in order to evaluate if the use of the calibration task data or the previous task data for each user may affect the predictive outcomes.

6.7.2. Physiological signals

Regarding the physiological signals, before generating the features to process, the data validity had to be evaluated. As it was seen in the experiment carried out in April 2016, due to the processing time of the AICARP application, some users’ signals were not collected in real time as the clock in the AICARP application was slower than the system one, adding a variable delay to their timestamps. In order to confirm that, the difference between the delay expected (100ms) between physiological registries and the real delay between the physiological registries recorded was plotted (Figure 41).

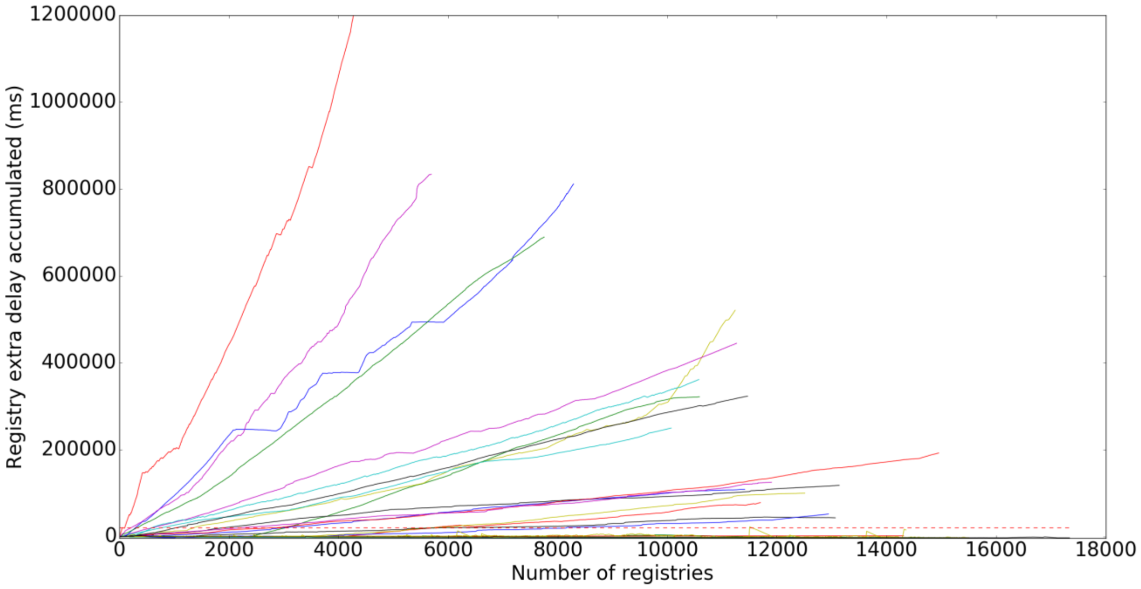


Figure 41. Accumulated delay in the physiological signals over time. Dotted line represents where the threshold for signal excluding was set.

The signals with an accumulated delay at the end of the experiment over 20000 ms were discarded. It can be seen in Figure 42 the final accumulated delay for each participant in experiment (and the threshold set). The maximum accumulated delay allowed threshold was chosen as, despite the maximum accumulated delay of the valid recordings at the end of the experiment was below 2000 ms, one of the recordings had a sudden increase of the accumulated delay during the last 8 seconds (80 registries collected at 10Hz) of the experiment (last 60 registries recorded for that participant) as can be seen in Figure 43.

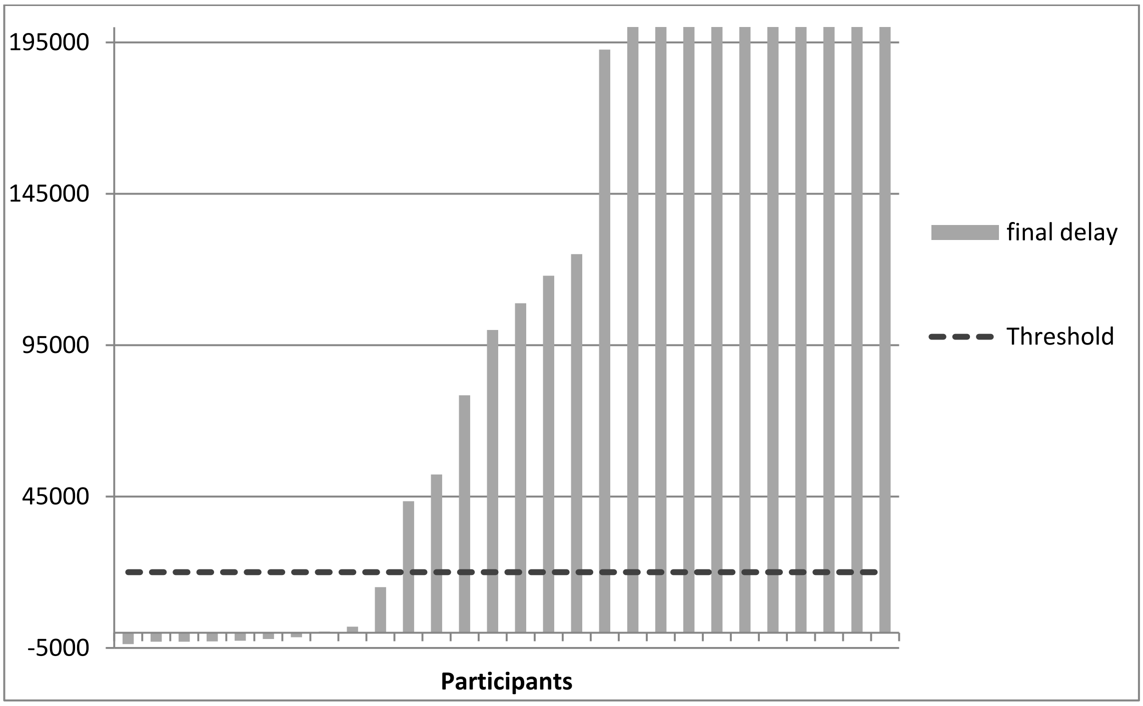


Figure 42. Final accumulated delay for each participant (experiment from April 2016).

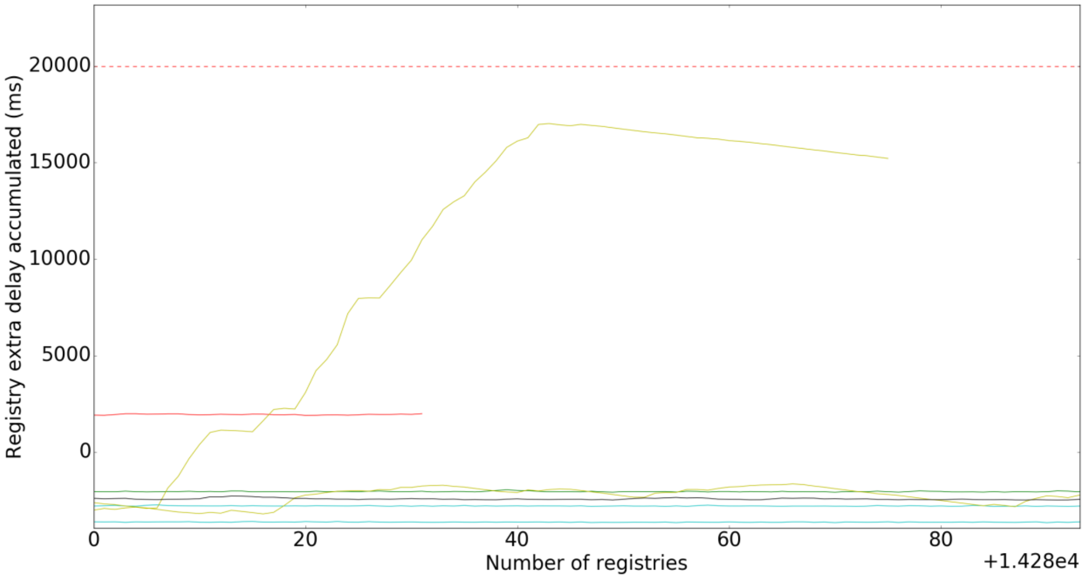


Figure 43. Detail of the accumulated delay in the recording of a physiological signal.

This problem in the recording was identified in the first experiment and reported by the author, and solved by Raúl Uría before the second experiment was held, so all the physiological recordings for the second experiment were correctly timestamped.

Once the defective recordings have been removed, the data was cleaned as done in stage 1 for removing noisy data. To do that, those registries containing a value out of a range considered “normal” were removed, and their value was set to an interpolated value taking into account the adjacent valid values. In case the recording of a signal contains more than a 30% of “invalid” values, that recording for that signal is to be discarded too. Then, heart rate variability is calculated following the code provided in [85], as it was seen to be used in many works presented in section 2.1.1 [14,51,125,148,184,186,242].

Also, another feature from the physiological signals is generated: the comparison between the values at the end of the task in contrast to the ones at the beginning. To do that, the whole signal is split into X slices (in our case X was set to 100), then, the first and last Y slices are discarded (in our case Y was set to 5) and finally, from the remaining slices, a given number of slices Z (in our case Z was set to 30) are kept at the end and at the beginning (discarding the “central” slices). The whole process can be depicted in Figure 44. Once the mean and standard deviation from the data contained in the remaining slices is calculated, the values are compared using the following formulas:

$$Sliced\ Mean\ Comparison = \frac{Mean(last\ slice\ values)}{Mean(initial\ slice\ values)} \qquad (6.3)$$

$$\begin{aligned}&Sliced\ Standard\ Deviation\ Comparison\\&\qquad = \frac{Standard\ deviation(last\ slice\ values)}{Standard\ deviation(initial\ slice\ values)}\end{aligned} \tag{6.4}$$

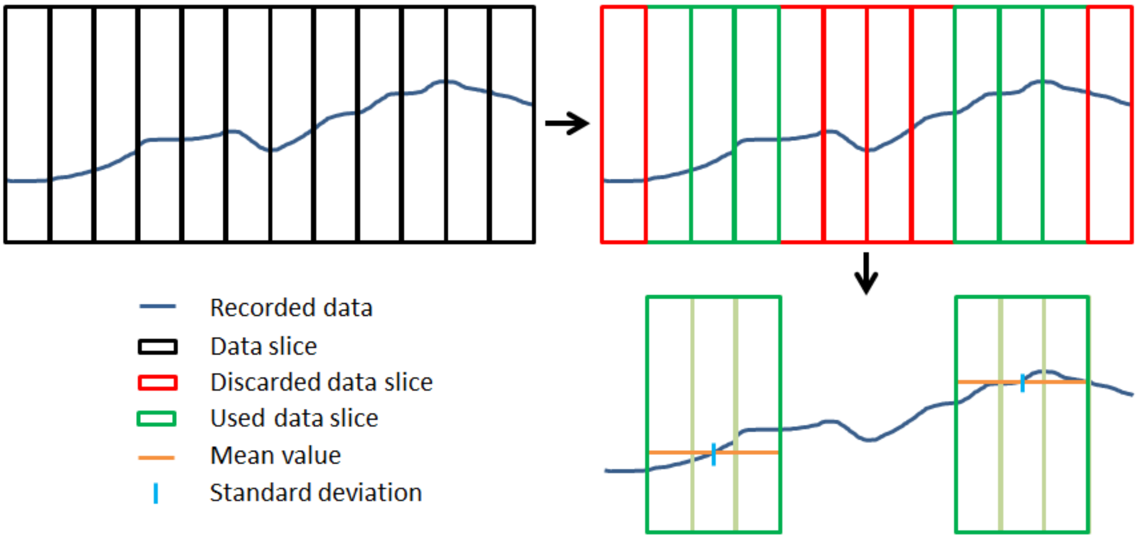


Figure 44. Sliced signal comparison feature generation process. In the case depicted in the figure, the signal is sliced into X=10 slices (in black), then, the Y=1 extreme slices (in red, at the extremes) are discarded and Z=3 slices (in green) are kept. The slices in the middle (also in red) are discarded. The mean (orange line) and standard deviation (blue line) is calculated from the remaining slices. At the end, these mean and standard deviation are compared.

After that, the features for each task and user are calculated from the raw values, calculating, for each user-task-signal combination the following features: mean value, standard deviation, maximum, minimum, number of registries, kurtosis, skew and sum.

Then, the original raw values are normalized using the initial baseline (following the same approach depicted in Figure 20 and Figure 21). And the features are calculated again, from the initial baseline normalized values, following the steps from the previous paragraph. After that, the original raw values are normalized again, this time using the previous task as baseline to normalize each task physiological values and the features are calculated as described in the previous paragraph.

6.7.3. Task Performance Model

A dataset reflecting the performance of the participants during the experiment was also generated. This way, since we are modeling affect in the educational domain, we can take advantage of performance model to improve the prediction when users' performance has an impact on their affective state (as in [137,171], where authors use the best trajectory in a task to evaluate the participant's performance). Bearing in mind that proposed tasks consisted in writing a short essay including certain selected words, the following features were included:

- Proportion of words used in the task compared with the mean number of words written by all the participants in that task (i.e. a comparison between how many words has the participant written in that task compared to others).

$$\frac{Number\ of\ words\ written\ by\ participant\ in\ task_i}{Mean(Number\ of\ words\ written\ by\ all\ participants\ in\ task_i)} \tag{6.5}$$

Proportion of proposed words used in the task (i.e. a score based on how many proposed words has the participant written).

$$\frac{Number\ of\ proposed\ words\ in\ task_i written\ by\ participant}{Number\ of\ proposed\ words\ in\ task_i} \tag{6.6}$$

Proportion of proposed words in the task used compared to the mean proportion of proposed words for that user in other tasks (i.e. a comparison between the score depicted in the previous point and the same score by the participant in the other tasks).

$$\frac{Number\ of\ proposed\ words\ in\ task_i written\ by\ participant}{Number\ of\ proposed\ words\ in\ task_{j\{j\in R|j\neq i\}}} \tag{6.7}$$

6.7.4. Sentiment analysis

As sentiment analysis was seen to be the most present data source in the best models generated, it was also included in this stage. Nevertheless, there are some big differences in the way the sentiment analysis is performed in this stage in contrast to how it was performed in the previous stage. Starting from the experimental design, as in this stage there was no explicit task asking the participants for their emotions explicitly. This time the goal was set to follow one open point identified in stage 1 (included in section 5.1) as the text to be analyzed is the result of the task itself, trying to avoid the inclusion of distracting tasks.

When analyzing the texts, the approach has also been updated since stage 1. As in this stage, the inclusion of certain terms (the proposed words to be included in the essay) may induce the topic of the essay, having an impact on the scores generated following the stage 1 sentiment analysis (e.g. one of the essays included proposed words related to felony and crime, and that topic may lead to negative scores). This time, the lexicon has been generated from the essays introduced by the participants, taking into account the scores of the SAM scores attached to the essays in order to provide an affective label to each term.

The sentiment analysis in this stage was performed as follows:

1. Essays were split into words
2. Stop words were removed
3. Words are stemmed using the Snowball Stemmer [181]
4. Rare words (those that appear in less than 3 documents) were also removed
5. For each essay, a bit vector is generated indicating which of the containing terms are included in that essay.

6. The SAM score given by the participant is set as the class attribute for each essay.
7. Using 10 fold cross validation and a decision tree learner, a sentiment analysis tag is given to each essay, which will be used as a feature in our system.

6.7.5. Labeling

As we have seen in stage 1, another important dataset to be generated was the one containing the label to be predicted. The affective model used, as in stage 1, is based on the two affective dimensions the participants were asked after each task (with values from 1 to 9 both of them): valence and arousal.

In related woks, these variables have been already discretized [254], but merged both valence and arousal into one single variable (with values: PVHA, PVLA, NVHA, NVLA and nVnA where P=positive, N=negative, H=high, L=low, n=neutral, V=valence, A=arousal) and did not give detailed data on the procedure used to carry out that discretization. In [42] different attribute thresholds are set to discretize the numerical class attribute (in this case is not affective state but cognitive demand), and the results of each threshold are evaluated in the experiment.

In stage 1, as seen in section 4.7.6 labels were discretized into positive or negative values. In this stage, the discretization process of the attribute to be predicted by the classification algorithms is another of the methodological variables that we consider, searching for a balance between fine grain modeling (the label given by participants) and prediction performance and simplicity (as the finer the grain, the worse the results might be and more complex the model). For this dataset, two versions (user-normalized and raw) were also generated aiming to design two different targets to evaluate: i) if the participant reflects a positive or negative absolute value in any of the emotional dimensions (raw version of the dataset) or ii) if the participant has gone through a positive or negative emotional transition from the beginning of the experiment (user-normalized version of the dataset).

- To model those positive or negative absolute numerical values (i.e., labels) given by the participants (raw version of the dataset), two different discretization approaches were applied:
    - The first discretization approach aims to draw a narrow neutral strip in the affective dimensions, considering positive (>5), neutral (5) and negative (<5) categories.
    - The second discretization approach aims to predict strongly positive (>6) and negative (<4) scores, with a wider neutral range (>4 and <6).

Thus, the second approach offers a high contrast between positive and negative states while the first one offers more data instances labelled with positive or negative values.

As participants reported on their affective state before doing the first essay task, we are also able to model the positive or negative emotional transitions from the beginning of the experiment compared to the end of each essay task by using the user-normalized version of the dataset.

- In order to evaluate different ways of modeling the changes produced two discretization approaches were also applied, as in the raw version of the dataset. The two followed approaches are:
  - A model including neutral value where the participant affective values are the same as in the calibration task, with positive (>1), neutral (1) and negative (<1) categories.
  - A model focused on detecting negative transitions with only negative (<1) and non-negative (≥1) categories.

These two different discretization approaches are depicted and highlighted in dash line boxes in Figure 45. The latter "imbalanced approach" is being explored here as has already been considered in our previous research on recommender systems [208], where the recommender system may only provide a recommendation when the learner is going through a negative affective state.

Additionally, to be used as a reference point, the label for the previous task was included in the dataset in order to provide information on how the participant is feeling right before the beginning of the task.

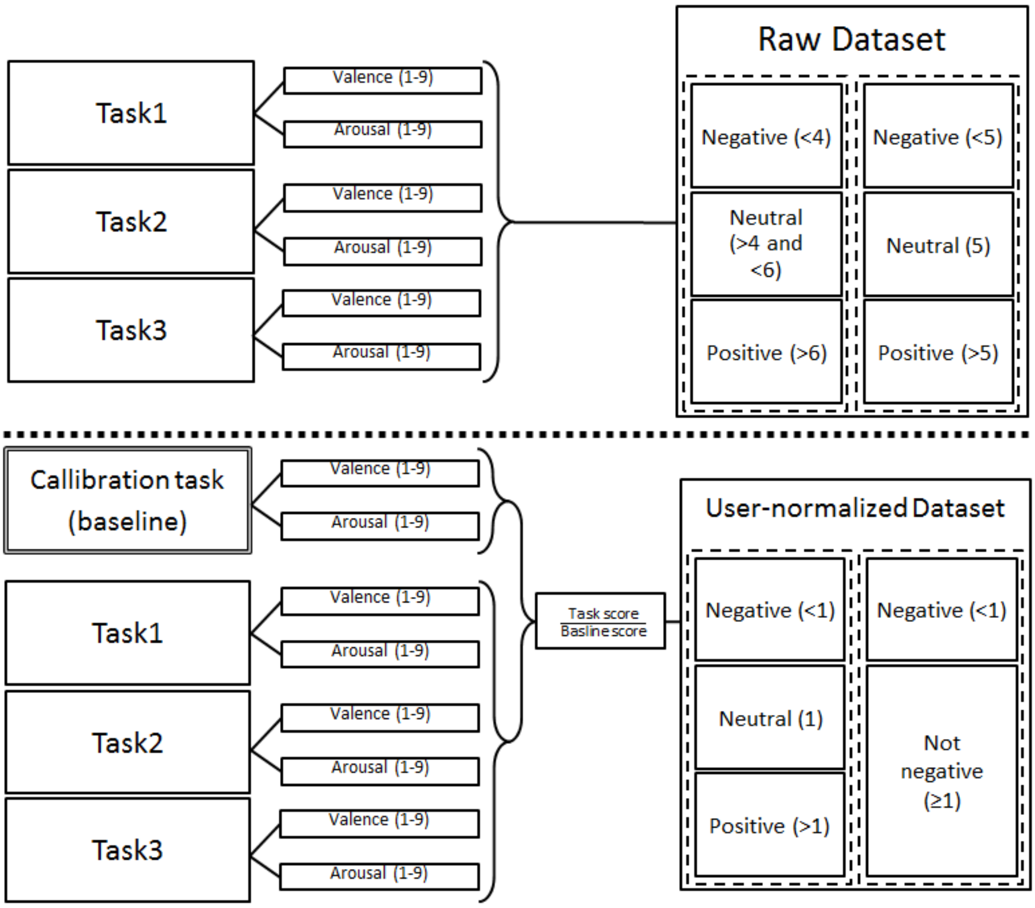


Figure 45. Different affective labeling discretization approaches used in stage 2. Each different approach proposed is included in a dash line box. Differential labels (bottom part) are generated by dividing labels given in regular tasks by labels given in the calibration task. Raw labels (top part) are generated from the labels given in regular tasks.

6.8. Data Processing

Once all the features have been generated, the models are to be generated. As done in stage 1 (and introduced in section 4.8), a workflow has been developed during the second half of 2017 in Knime in order to generate all the models required for our research. These predictive models are to be generated from the combinations of all the datasets already presented using data mining techniques. A total of 31105 models have been generated in this second stage. Table 27 includes the variables that have been taken into account in the model generation (discussed in section 6.2).

Variable identifier	Variable name	Description	Possible values	Dependencies
Variable MV1	Target attribute	The attribute to be used to label the data (and the attribute to be predicted	Valence	
			Arousal	
Variable MV2	Clustering	Clustering technique being used with the dataset.	None: the features are being used as input to generate the model	
			Cascade simple K-means: K-means clustering using [46] as a criteria for selecting the best K.	
			EM clustering: clustering using the expectation maximization technique.	
Variable MV3	2-step classification	The 2-step classification approach based on predicting emotion vs. neutral state (first stage) and detect the the emotion (if in the first stage that registry was labeled as not neutral)	No: The classification is done directly from the class attribute.	
			Yes: The class attribute is renamed to “neutral” & “not-neutral” according to each registry label, then, a first classification is done to predict that attribute. Then, those registries classified as “not neutral” are used to predict if the class attribute is positive or negative.	

Variable identifier	Variable name	Description	Possible values	Dependencies
Variable MV4	Normalization	The normalization process to be used on the data (i.e. mouse, keyboard, physiological signals and class attribute label). In case the normalization is performed, it is done following the formula (6.2).	No: The data to be used is left as collected.	When no normalization process is performed, the features described in 6.7.1.a are not used.
			Using a fixed baseline: The data is normalized using as reference values those recorded in the calibration task (except the physiological signals, see variable 5 in this table)	
			Using a dynamic baseline: The data is normalized using as reference values those recorded during the previous task.	
Variable MV5	Baseline used for physiological signals	As done in stage 1, two baselines are recorded for the physiological signals (one prior to the beginning of the experiment and another one after its end). Nevertheless, during the calibration task, physiological signals were also recorded.	Pre-experiment baseline: Using the values recorded from the baseline recorded before the participant starts any task.	Variable MV4 in this table (value: Using a fixed baseline)
			Post-experiment baseline: using the values recorded from the baseline recorded at the end of the experiment.	Variable MV4 in this table (value: Using a fixed baseline)
			Pre & post-experiment baseline combination: using the values recorded in both physiological baselines, at the beginning and the end of the experiment.	Variable MV4 in this table (value: Using a fixed baseline)
			Calibration task baseline: using the values recorded during the calibration task.	Variable MV4 in this table (value: Using a fixed baseline)
Variable MV6	Feature selection technique	The technique used to perform feature selection in order to	None	
			Forward feature selection	

Variable identifier	Variable name	Description	Possible values	Dependencies
		reduce the dimensionality of the dataset	Principal Component Analysis	
Variable MV7	Class balancing technique	The technique used to balance the dataset according to the class attribute	None	
			SMOTE oversampling	
			Equal Size Sampling	
			Both	
Variable MV8	Discretization approach	The data discretization approach used with the class attribute.	Negative (1-3), Neutral (4-6), Positive(7-9)	Variable MV4: no normalization
			Negative (1-4),Neutral (5), Positive (6-9)	Variable MV4: no normalization
			Negative(<1), Not negative (≥1)	Variable MV4: normalization (both fixed or dynamic baseline)
			Negative (<1), Neutral (=1), Positive (>1)	Variable MV4: valor normalization (both fixed or dynamic baseline)
Variable MV9	Algorithm	The algorithm used to generate the prediction model	J48	
			Naïve Bayes	
			Random Forests	
			SMO	
			Bagging	
			Bayes Net	

Table 27. Methodological variables taken into account in the model generation process in stage 2.

We can see how there have been many methodological variables taken into account in this stage. Here we are going to discuss the implementation of the workflow generated:

Initially, all the data is loaded. After that, the data from the data sources is normalized depending on variable 4 (and in case of physiological signals, also depends on variable 5). In addition, one of the class attributes is selected to be used as label to be predicted (depending on variable 1) and discretized according to variable 8. The unused emotional dimension is discarded. After that, some features are filtered out using the following criteria: i) remove columns or rows with more than 40% of missing values; ii) remove highly correlated features (i.e. a correlation index over 0.7). In addition, performance features as well as sentiment analysis scores are calculated. The affective score of the previous task is also included in the dataset. Once all the features have been merged into a single dataset, depending on the variables 6 & 7, some preprocessing techniques might be applied. Then, depending on the technique pointed by variable 2, several clustering models are generated: i) clustering is performed with all the features available; ii) another clustering model is generated from keyboard features; iii) another clustering model is generated from mouse features; iv) another clustering model is generated from physiological features; v) another clustering model is generated from the remaining features (i.e. performance, sentiment analysis score and label from previous task); vi) another clustering model is generated from the features depicted in section 6.7.1.a (if there is a normalization process performed depending on variable 4).

Depending on the clustering performed (set by variable 2) we will have at this point a dataset consisting on different clusters (if a clustering has been performed) or the dataset previous to the clustering process (if variable 2 was set to “no clustering”). At this point, in case 2-step classification has to be performed (set by variable 3), an initial model is generated to discard those registries reflecting a neutral affective state, keeping the remaining ones. After that, the current dataset is used to perform 6 different predictive models (based on the 6s algorithms depicted by variable 9). Figure 46 depicts the data processing workflow generated. There we can see the way the data is processed and where are the different methodological issues introduced addressed. The different methodological variables depicted in Table 27 are pointed in the figure in the node where the corresponding data processing for that variable is performed. The results of the models generated are to be discussed in next section.

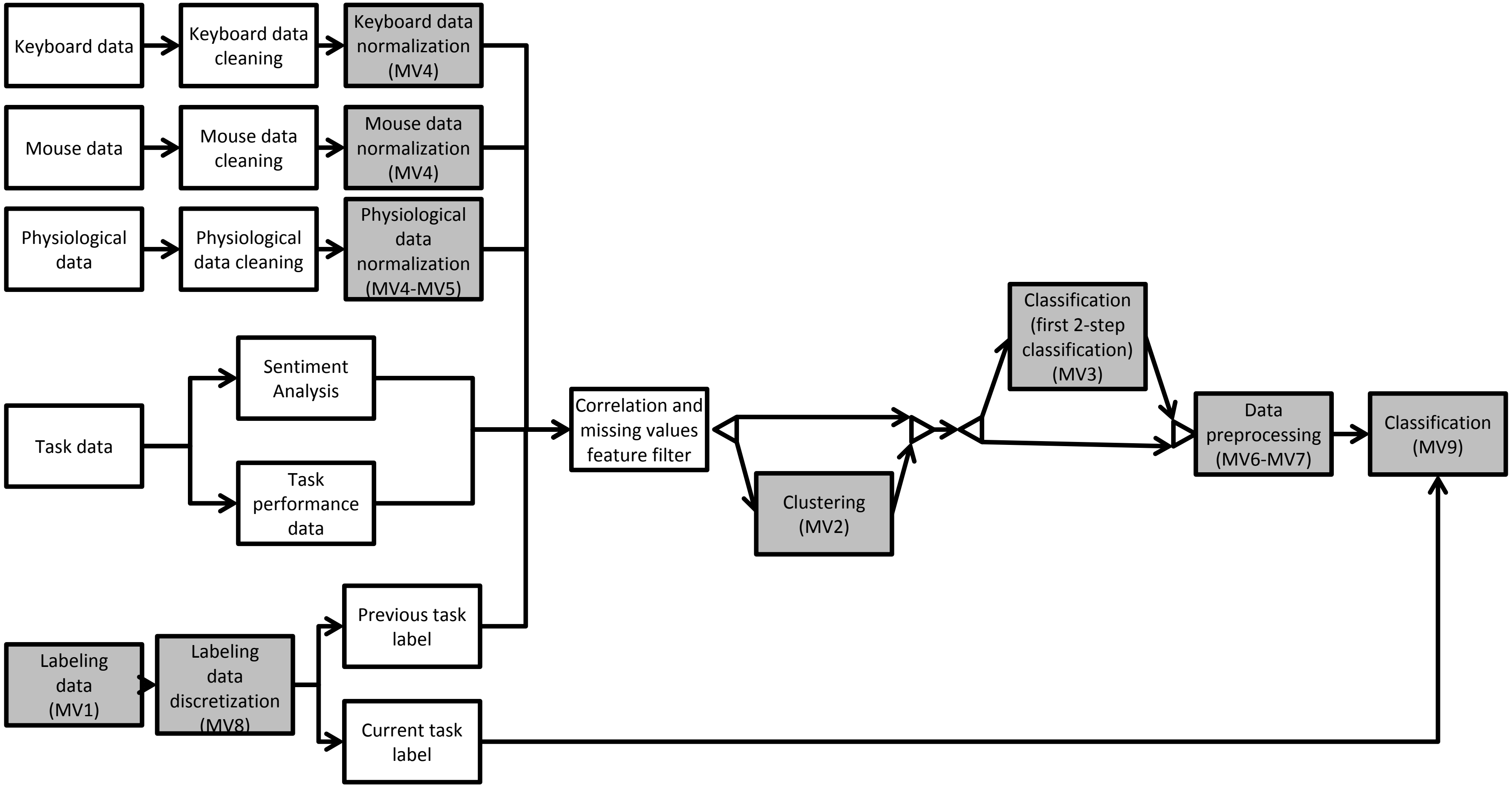

Keyboard data
Keyboard data cleaning
Keyboard data normalization (MV4)
Mouse data
Mouse data cleaning
Mouse data normalization (MV4)
Physiological data
Physiological data cleaning
Physiological data normalization (MV4-MV5)
Sentiment Analysis
Task data
Task performance data
Correlation and missing values feature filter
Clustering (MV2)
Classification (first 2-step classification) (MV3)
Data preprocessing (MV6-MV7)
Classification (MV9)
Previous task label
Labeling data (MV1)
Labeling data discretization (MV8)
Current task label

Figure 46. Data processing workflow. Different data sources data is cleaned and normalized separately and then put together with sentiment analysis, task performance and the previous task label. After that, correlated features and features with missing values over a given threshold (50%) are filtered out. Then, in case clustering variable is being evaluated, the data is clustered depending on the data source. After that, in case the 2-step classification approach is being evaluated, the first classification is performed in order to detect neutral or not-neutral states, letting only those not neutral states go to the next classification step. Finally, the different data preprocessing techniques evaluated are applied and the models are generated. Nodes with grey background are those where the methodological variables (MV) described in Table 27 are processed.

6.9. **Results**

With all the variables described in the previous section, 31105 models were generated, all of them applying 10-fold cross validation. In this section we are about to show the best results according to the models generated from the datasets created taking into account each one of the possible values of the methodological variables introduced in Table 27.

In the following subsections, the results of the evaluation of the impact of the methodological variables depicted in Table 27 are shown. To do that, the best model (i.e. the model with the highest value of the indicator depicted in formula (4.6)) for each value of the variable analyzed is going to be depicted in a figure by mean of its accuracy, Cohen's Kappa and accuracy improvement (i.e. the difference between the accuracy of the model and the dataset majority class). After that, another figure is going to depict the average accuracy, Cohen's Kappa and accuracy improvement of all the models built with each value of the methodological variable evaluated. The sets of results for the best models are to be represented connected by a line in order to ease the comparison between the values of the different models. The results from the top 100 models for valence and for arousal can be found in Appendix V (section 13.5).

6.9.1. Results for different normalization baseline used

The evaluation of this variable aims to match the first objective defined for this stage (**O2.1** in in section 1.4): *Evaluate the impact of user centered normalization in across subject experimental approaches.*

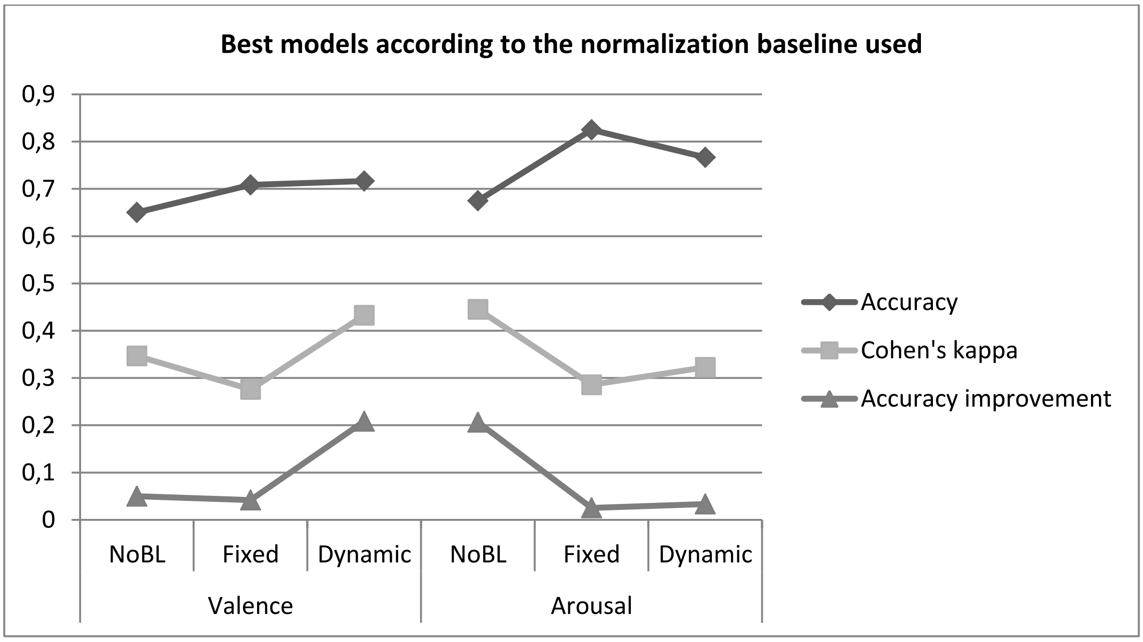


Figure 47. Best models according to the normalization baseline used.

These results can also be seen in section 6.9.1.As we can see in this first graph, when predicting the valence, dynamic baseline seems to provide the best predicting results. In contrast, when predicting the arousal dimension, not using any normalization seems to provide the best models. This is something we are going to see in the following figures, as the top models generated from the discussed variables are going to provide its best results in those datasets predicting valence using normalization from a dynamic baseline and those datasets predicting arousal with no normalization applied to the dataset.

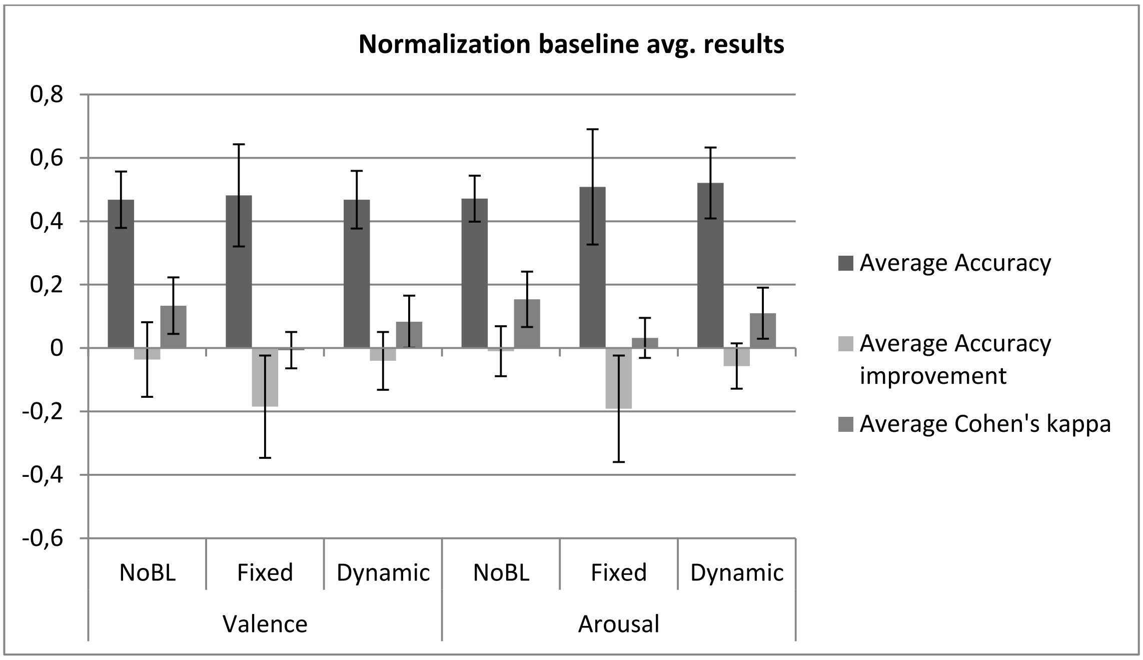


Figure 48. Avg. results from all the models aggregated by the normalization baseline used

When evaluating the avg. Values of all the models generated, we can see that fixed baseline provides the worst average kappa and accuracy improvement as well as a higher standard deviation in the accuracy results. No baseline and dynamic baseline seem to offer similar results (with the no baseline models showing a slightly better average kappa).

In conclusion, the use of a dynamic baseline seems to be a good approach. When predicting affective valence, the best model based on a dataset generated from a dynamic baseline has provided better results than any of the alternatives proposed. When evaluating the aggregated results of all the models generated with each approach, dynamic baseline normalized models seem to offer similar results to the ones that have not been normalized.

6.9.2. Results from Clustering-based approach

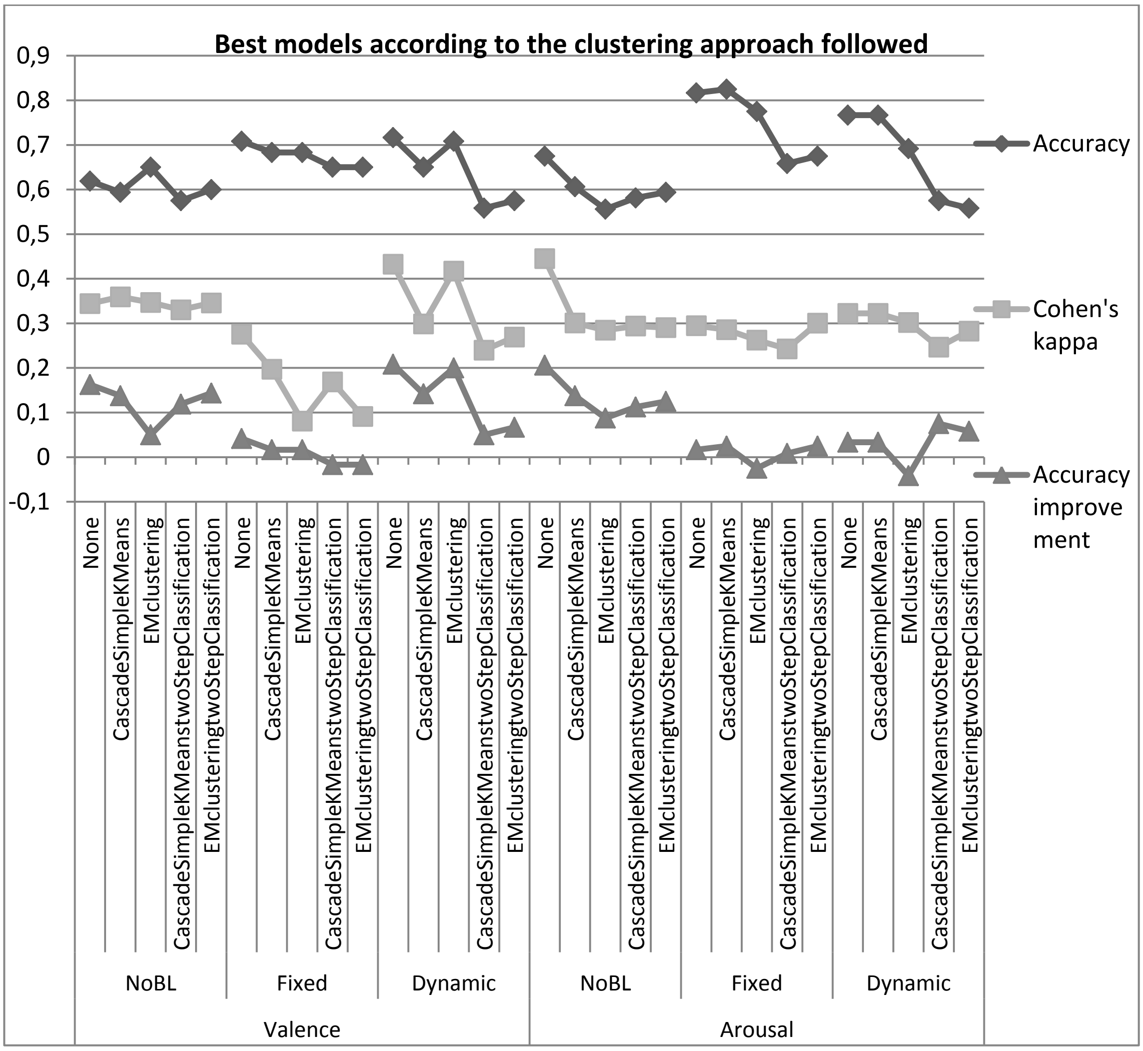


Figure 49. Best models according to the clustering approach followed

If we analyze the results from the best models from the different clustering approaches carried out, we see that no clustering provides always the best (or close to the best) results. When any clustering technique is used, despite 2-step classification has been performed, similar results are obtained. Only three models provide Cohen's Kappa values over 0.4: two of them when the dynamic baseline is used to predict valence (with no clustering technique and with Expectation Maximization clustering technique) and the other one, predicting arousal with no baseline and no clustering technique. These three models are also the ones offering the best accuracy improvement values of all the models (providing accuracy close to a 20% better than the majority class ratio).

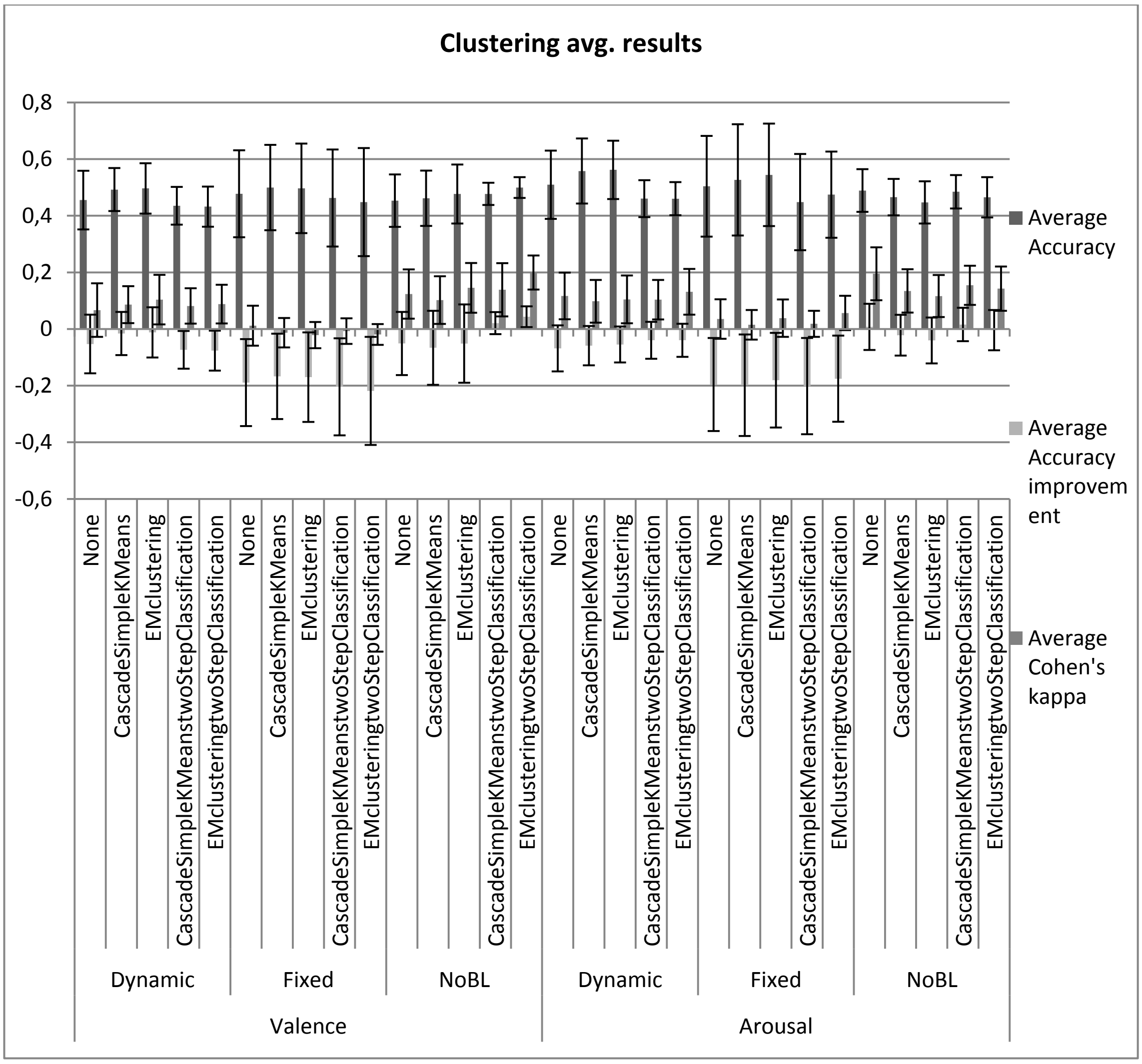


Figure 50. Avg. results from all the models aggregated by the clustering approach followed

As we can see evaluating the use of clustering techniques, the best average kappa value is obtained from using the EM clustering technique combined with the 2-step classification approach with no data normalization on predicting valence. It seems that EM clustering (in some cases combined with 2-step classification) offers best (or close to best) average kappa values over the rest of the options more often than any other approach, with the exception of predicting arousal with no data normalization, where no using any clustering technique offers the best kappa values.

6.9.3. Results from 2-step classification approach

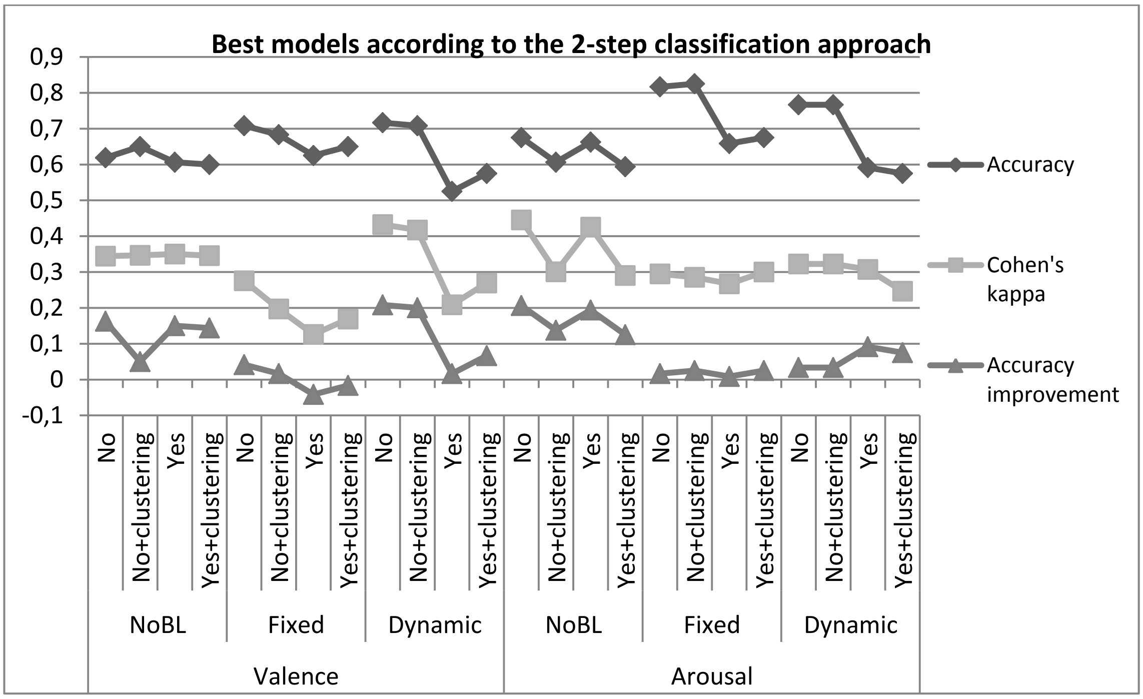


Figure 51. Best models according to the use of the 2-step classification approach

As we could see in section 6.9.1 (Figure 47), best results appear using no normalization when predicting arousal. When predicting both valence and arousal from a normalized dataset, 2-step classification provides worse top models than not using 2-step classification. When not using any normalization on valence predicting, using the 2-step classification approach provides similar results to the ones obtained with not using it.

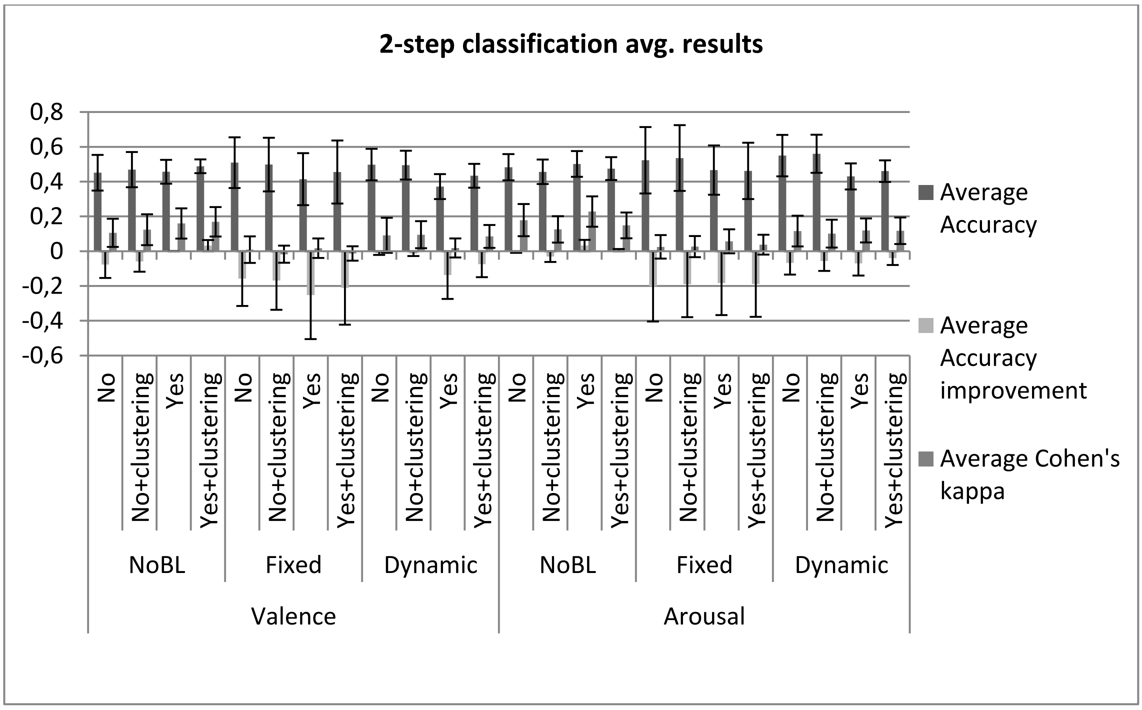


Figure 52. Avg. results from all the models aggregated by the use of the 2-step classification approach

Evaluating the aggregated results from the models grouped according to the application of the 2-step classification proposed approach, we can see how the application of that 2-step classification approach provides similar or better kappa values in most of the cases, with the exception of predicting valence with the dynamic baseline normalization.

6.9.4. Results from class-attribute discretization approach

The results shown in this section aim to carry out the evaluation introduced by the objective **O2.3** described in section 1.4: *Evaluate different data discretization approaches on affective numerical labels.*

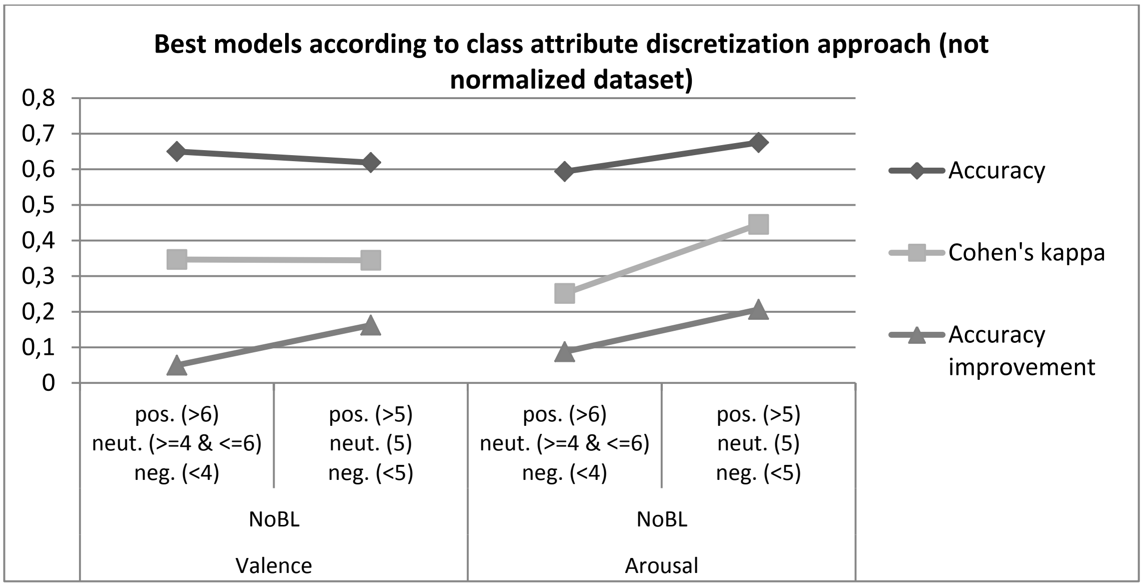


Figure 53. Best models according to class attribute discretization approach (models built from the not normalized dataset)

Looking at the discretization approaches used in the not normalized dataset, we can see that top best models perform better when using a narrow range for the neutral affective states (especially when predicting the arousal dimension).

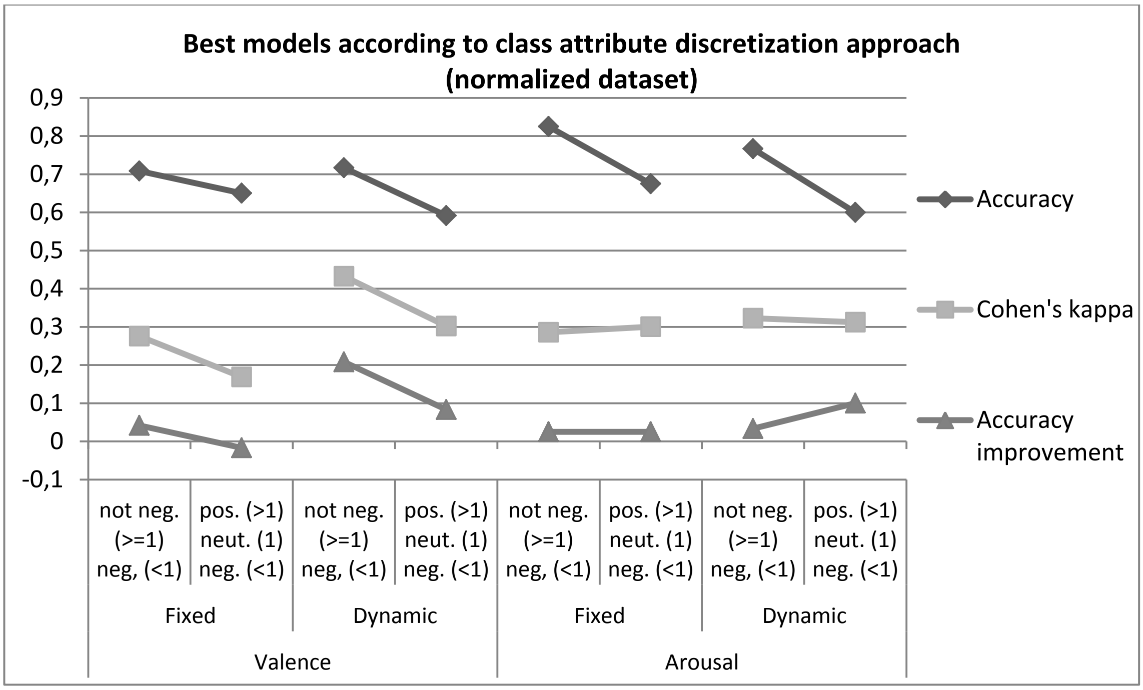


Figure 54. Best models according to class attribute discretization approach (models built from the normalized dataset)

When evaluating the discretization approach followed when the dataset has been normalized, we can see how, in case of predicting valence, using the 2-class approach provides a better top model (with higher accuracy, kappa and accuracy improvement) than using the 3-class approach. When predicting arousal, both discretization approaches seem to provide a similar top model (although the 3-class approach seems to provide lower accuracy rates).

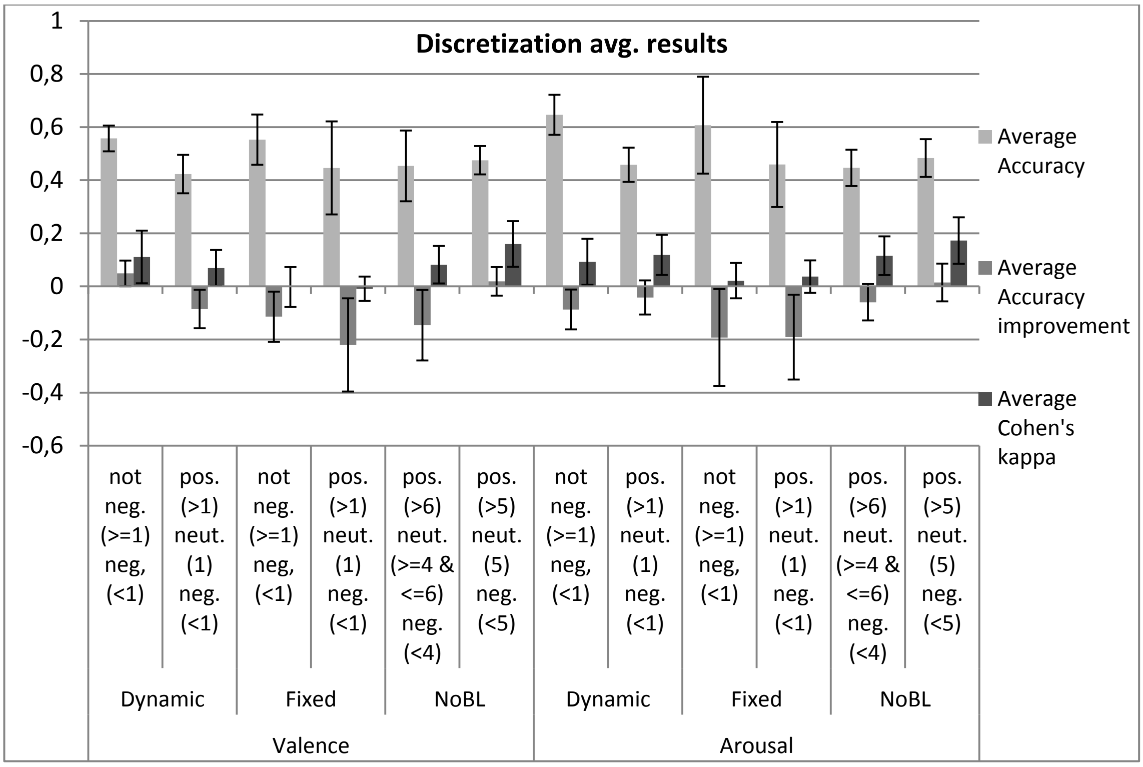


Figure 55. Avg. results from all the models aggregated by class attribute discretization approach

As we can see, in not-normalized datasets aggregated results, best Cohen’s kappa and accuracy improvement values are provided when using a narrow range for the neutral affective states. When evaluating the aggregated results for the models generated from normalized datasets, the 2-class approach provides higher accuracy and accuracy improvement values.

6.9.5. Results from class balancing technique used

The results shown in this section aim to show the results obtained from the research driven by the objective **O2.2** described in section 1.4: *Evaluate different preprocessing techniques on the data collected and their impact on the results*.

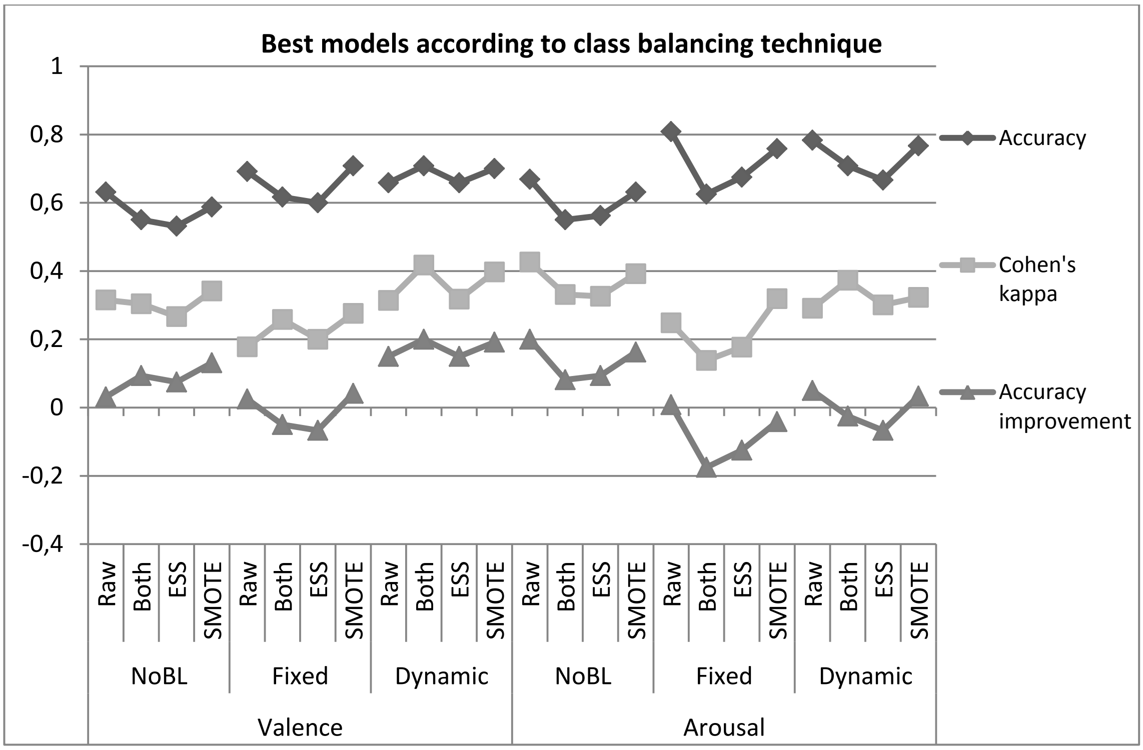


Figure 56. Best models according to class balancing technique

If we look at the class balancing techniques used, we can see how, again, the best models are in the dynamic-baseline normalized valence dataset and the non-normalized arousal dataset. In the valence prediction models, we can see how the use of SMOTE technique provides the best results. In the arousal prediction models, SMOTE and no-class balanced datasets seems to provide similar top models (over ESS and both techniques datasets).

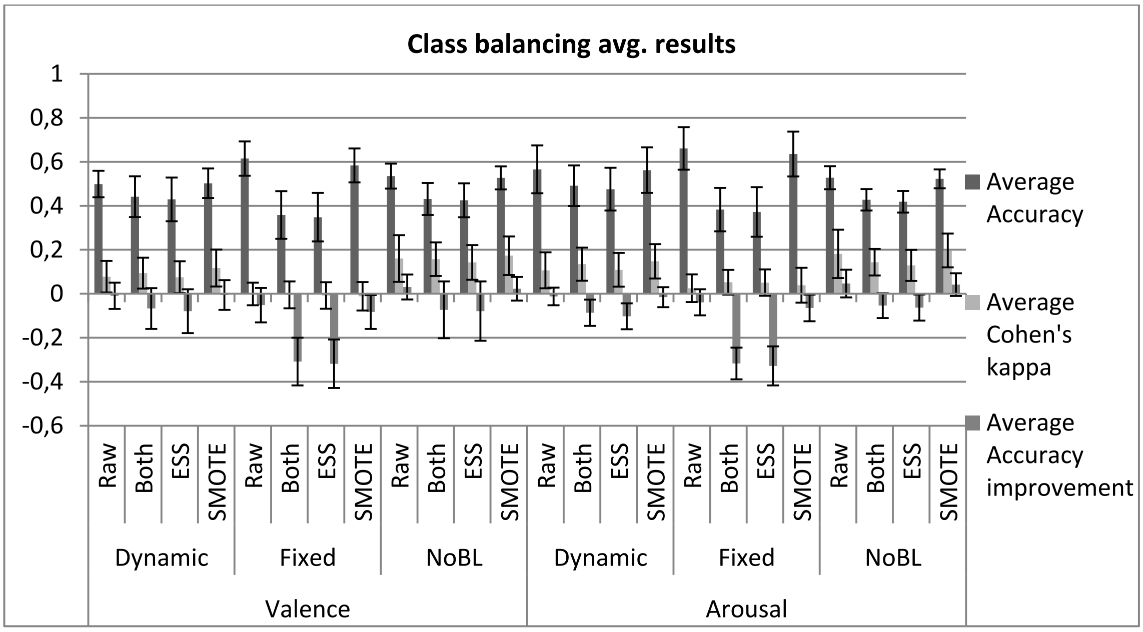


Figure 57. Avg. results from all the models aggregated by class balancing technique

We can see in Figure 57 how results get worse when Equal Size Sampling is being applied (with or without SMOTE). The average kappa values provided by the models built using the SMOTE technique seems to be the best ones (although sometimes the raw dataset provides similar results).

6.9.6. Results from feature selection technique used

The results shown in this section aim to show the results obtained from the research driven by the objective **O2.2** described in section 1.4: *Evaluate different preprocessing techniques on the data collected and their impact on the results*.

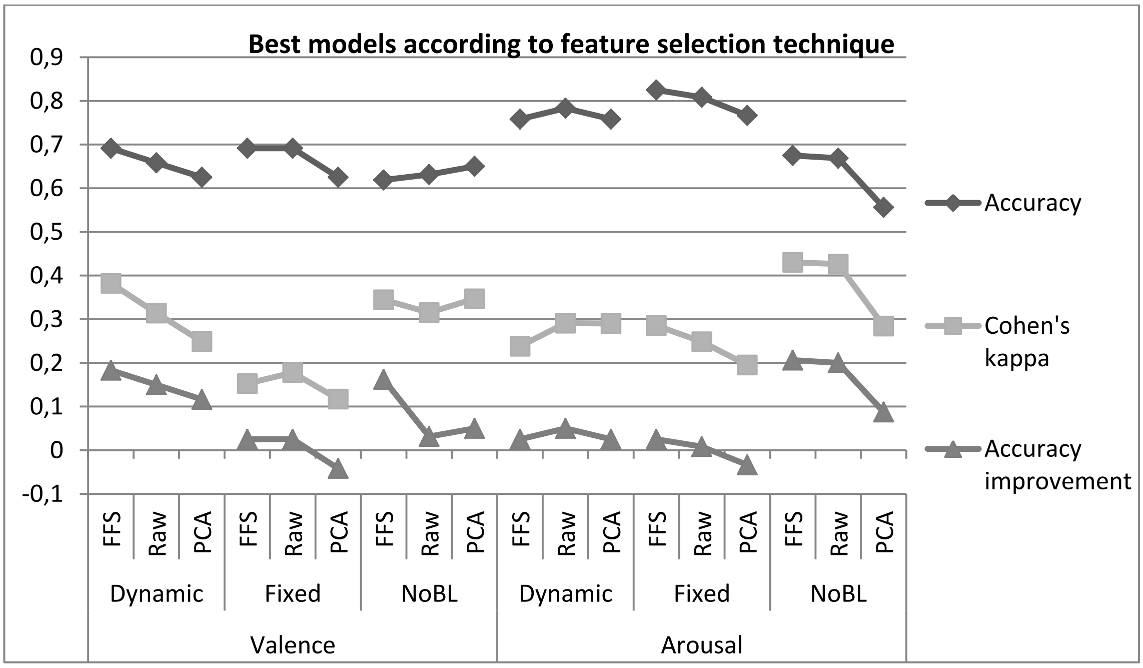


Figure 58. Best models according to feature selection technique

When analyzing the feature selection techniques used in the different datasets generated, it seems that forward-feature selection technique provides best results in most cases. Not using any technique provides similar or slightly worse results than using FFS, while using PCA seems to provide the worst results in most of the datasets proposed in the figure.

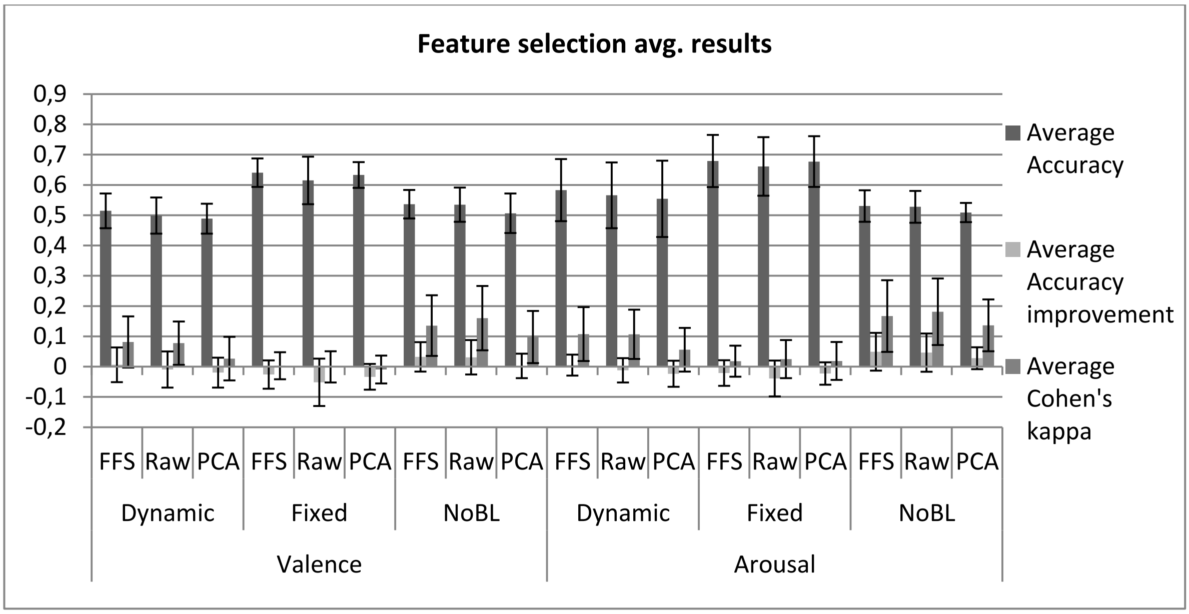


Figure 59. Avg. results from all the models aggregated by feature selection technique

Evaluating the avg. Values of the models according to the feature selection technique used, we can see how using Forward Feature Selection provides better or similar accuracies to the ones provided by the models built from the raw dataset (with similar kappa values). Models built from the PCA datasets provide similar or slightly worse values (especially kappa values).

6.9.7. Results from algorithm used

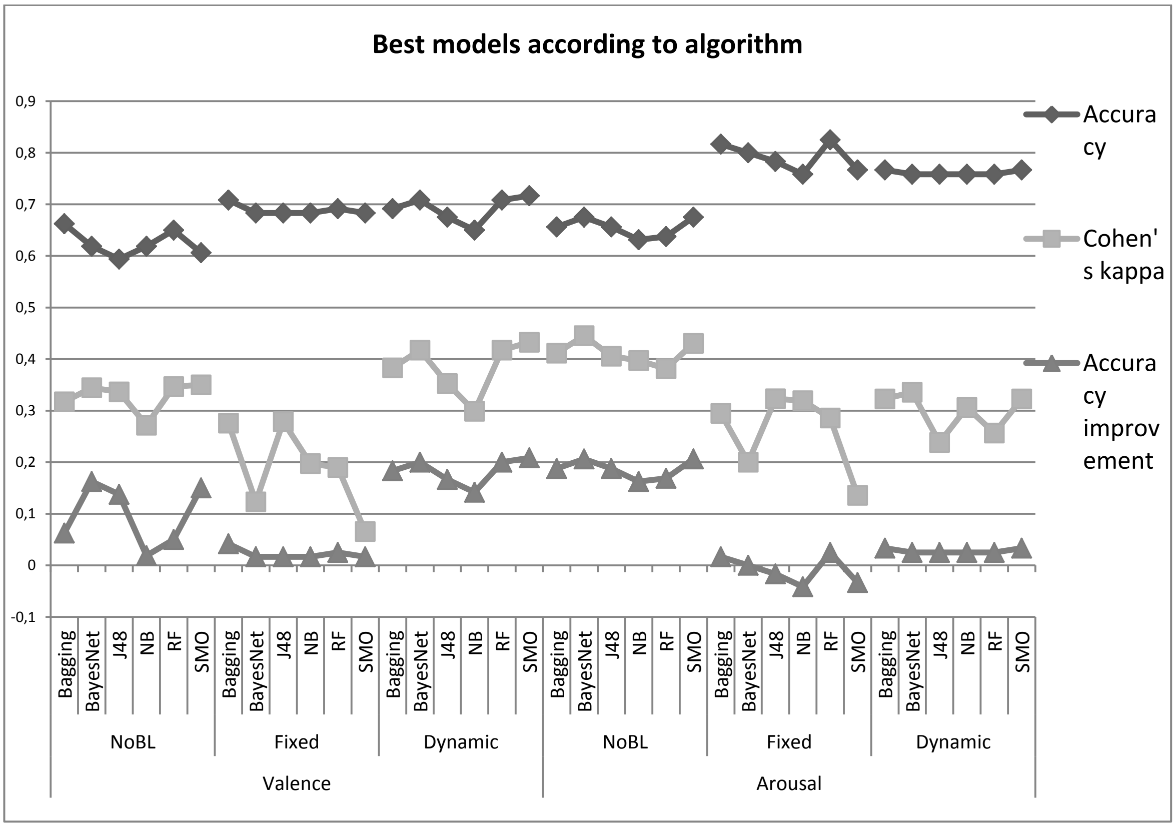


Figure 60. Best models according to machine learning algorithmused

Analyzing the different algorithms used, we can see how Bagging provides in many cases, best or similar to the best results. Something similar happens to bayesnet and SMO, which behave in a similar way, providing good results in most cases, but showing poor results when the normalization is performed with a fixed baseline. J48, NB and RF seem to show a different performances depending on the dataset used.

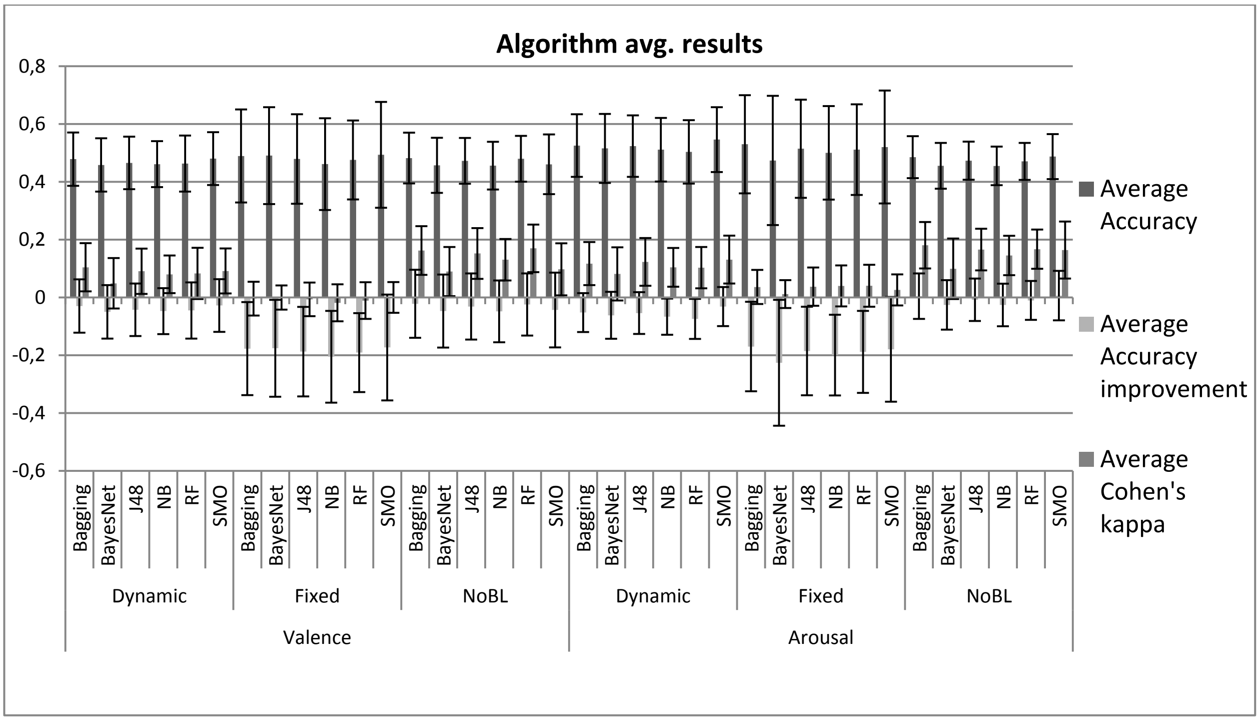


Figure 61. Avg. results from all the models aggregated by machine learning algorithm used

When evaluating the aggregated results according to the algorithm used, none of them seem to provide a clear advantage over the others.

6.9.8. Results from baseline used for the physiological signal

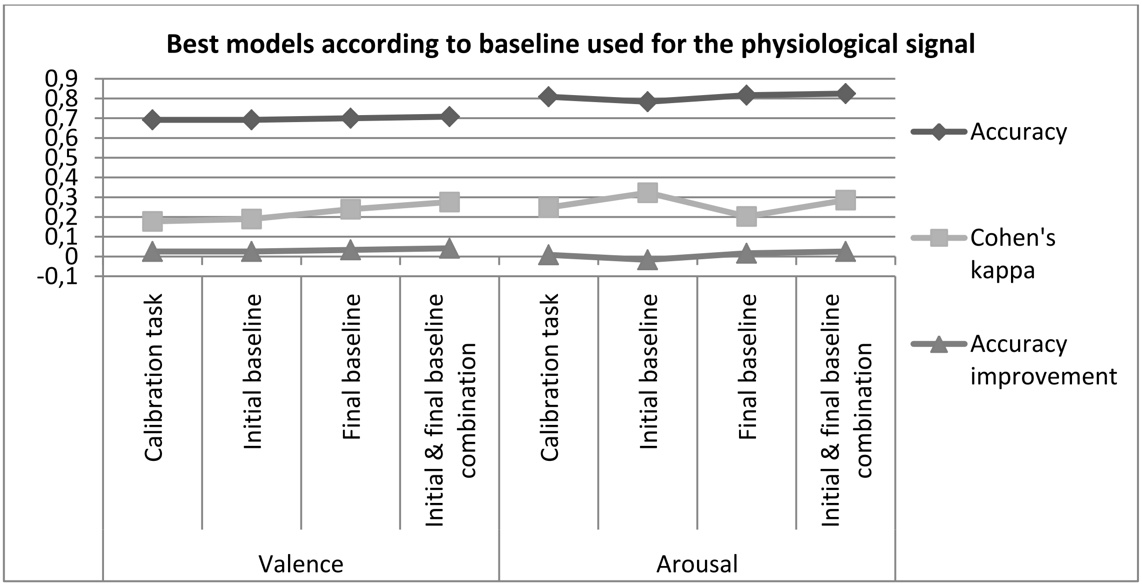


Figure 62. Best models according to baseline used for the physiological signal used

Analyzing the physiological baseline used (when the normalization is performed from a fixed baseline), we can see how the combination of initial and final baselines provide the best models both predicting valence or arousal dimensions. Using the physiological signals recorded during the calibration task seems to provide the worst top model results.

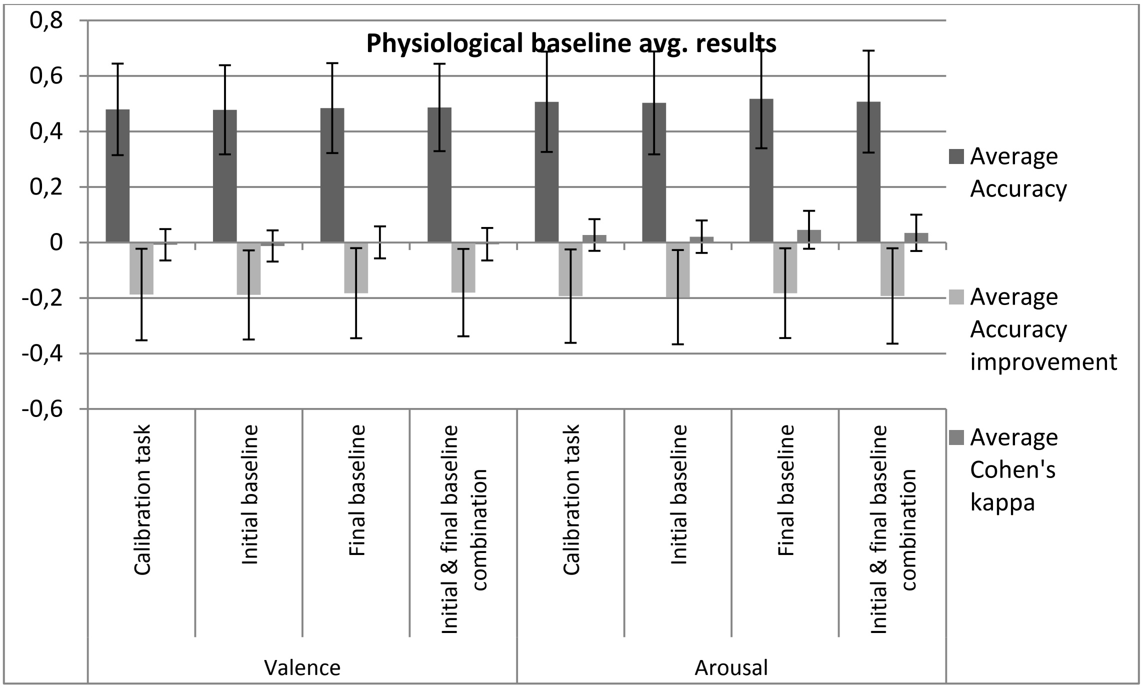


Figure 63. Avg. results from all the models aggregated by physiological signal used

When evaluating the baseline to be used for the physiological normalization, the different approaches seem to provide similar results. The baseline recorded after the experiment might provide very slightly better kappa values.

6.9.9. Results analysis

The 50 best models (according to the best values given by Formula (4.6)) were chosen for each one of the two dimensions to predict. Figure 64 represents the aggregated results from those top 50 models:

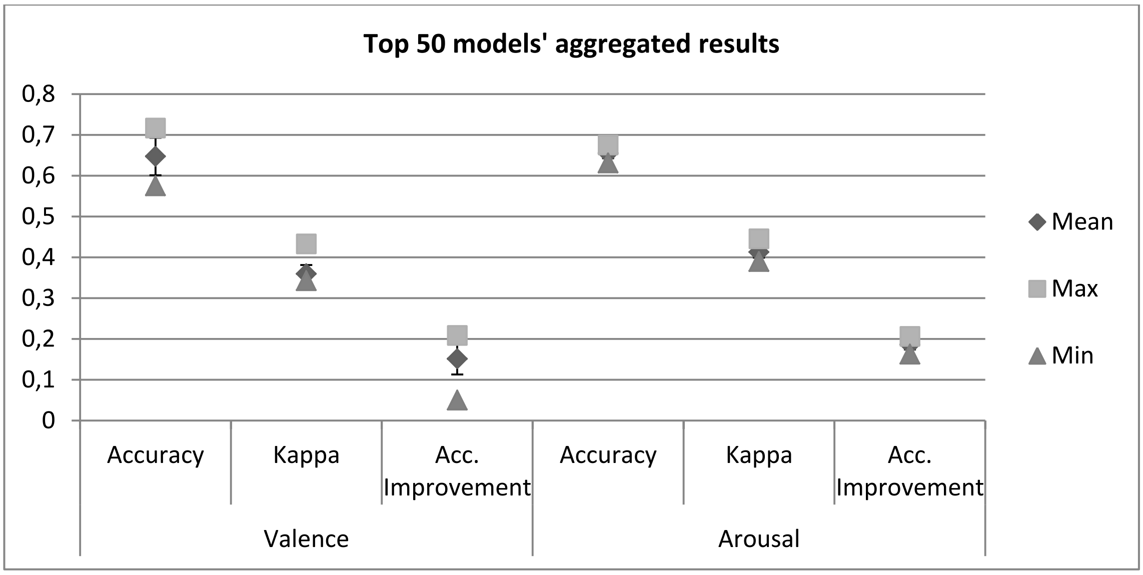


Figure 64. Accuracy, Cohen's Kappa and Accuracy improvement (difference from base rate) from the top 50 models generated for predicting valence and arousal

As we can see, when predicting valence or arousal, similar accuracy levels are obtained from the top algorithms. When predicting valence, we can see how kappa values are slightly lower to the arousal ones, as also happens with the accuracy improvement.

Comparing our best results with similar related works, we can find low kappa values in top results from some studies: in [28], where best results distinguishing engagement from boredom offer a 0.374 kappa value (87.0% accuracy rate) and best results distinguishing three emotions offer a 0.171 kappa value (56.3% accuracy rate). Top classifiers from [76], predicting single emotions, provide accuracies from 76,3% (Kappa=0,55) to 93,8% (Kappa=0,55). The models from [76] were built using a reduced data set (reduced initially by removing the neutral category and further during under-sampling). Accuracy values presented in [122] vary depending on the emotion predicted and the algorithms used, showing values from 47.37% to 81.25% (mean=62.47%, std. dev.=8.67). In case of [171] accuracy rates range from 91,96% to 94,61% when predicting confusion and data about a known target is used. When no information about the target is used, the accuracy rates go from 82,38% to 84,47%. Bearing in mind the critical issues involved in our experience, reported related work provides similar results to the ones we obtained. In particular, when there is a small number of data, large number of features and elusiveness of the target outcome, which, as it will be pointed out in the following section, responds to the ultimate goal and circumstances involved. In particular, our work uses a discretized dimensional approach, while most similar works use a closed set of emotions, which can produce different results depending on the emotion predicted [76,122].

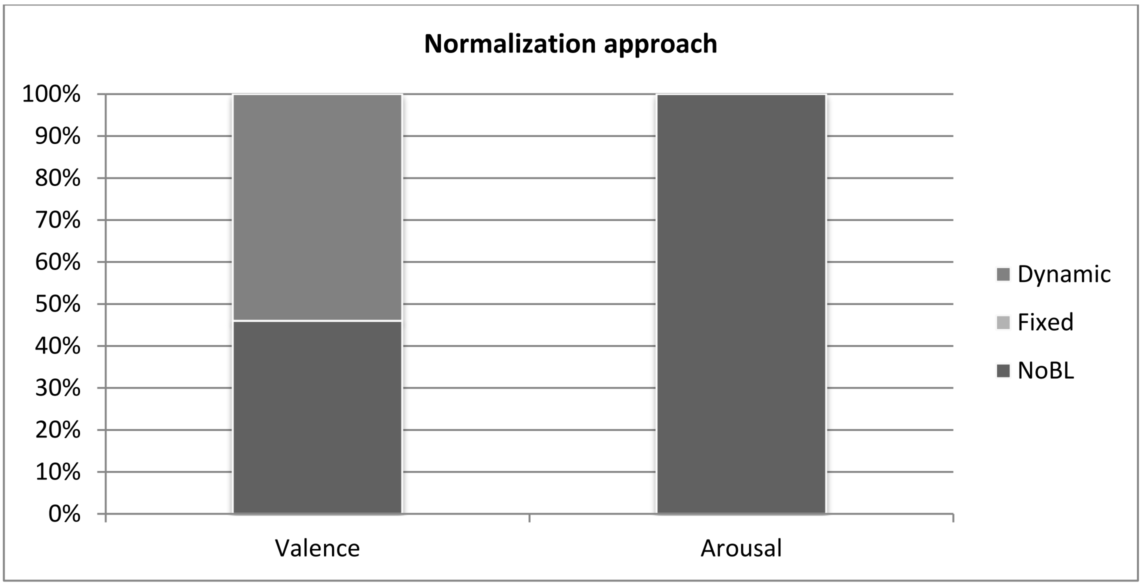


Figure 65. Normalization approaches used in the top 50 models for each emotional dimension.

In Figure 65 it can be seen how the top valence prediction models are obtained from not-normalized or normalized using the dynamic baseline, while the 100% of the 50 top models for predicting arousal is achieved from un-normalized datasets. This conclusion partially validates the hypothesis (**H2**) introduced in section 1.4 that claimed that: *In real-world learning scenario based inter-subject experiments, the use of a reference state to normalize each user interaction related data provides more robust models when detecting affective features to enrich learner modeling in educational contexts.* As we have seen in section 6.9.1, the models calculated from the dynamic baseline based normalization, provides a more robust best model as provides better accuracy, accuracy improvement and Cohen's Kappa than the best model built from the not normalized dataset.

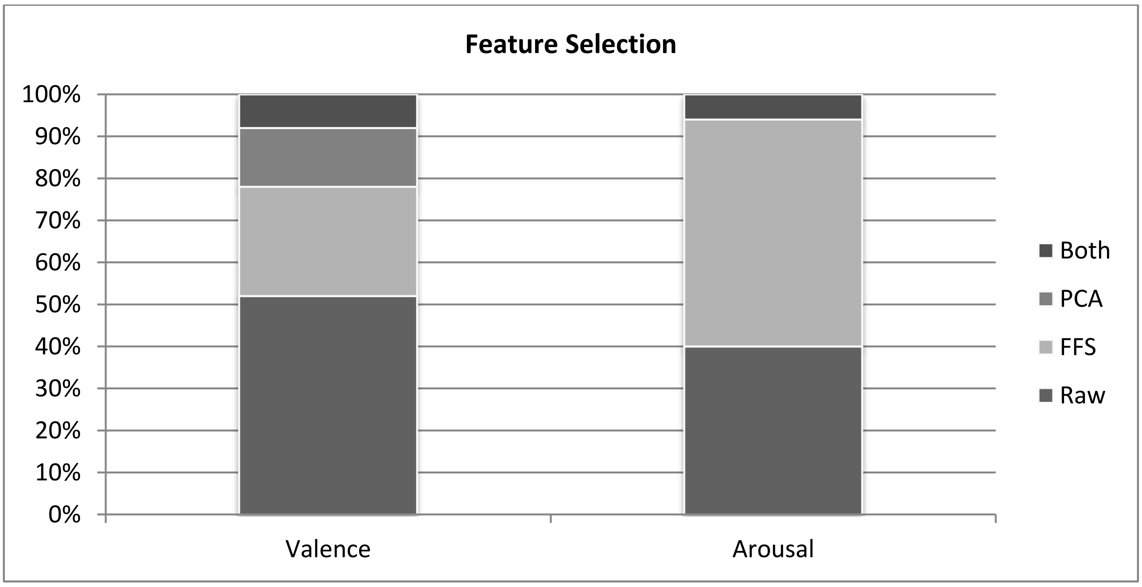


Figure 66. Feature selection techniques used in the top 5 models for each emotional dimension.

Figure 66 shows the use of feature selection techniques in the top 50 models generated for each emotional variable. FFS is used in 54% of the top models predicting arousal, while it is only used in the 26% of the models predicting valence. PCA is used only in a 22% of the valence models and 0% of the arousal models. Both techniques provide the best results in an 8% (valence models) and a 6% (arousal models) of the top models generated.

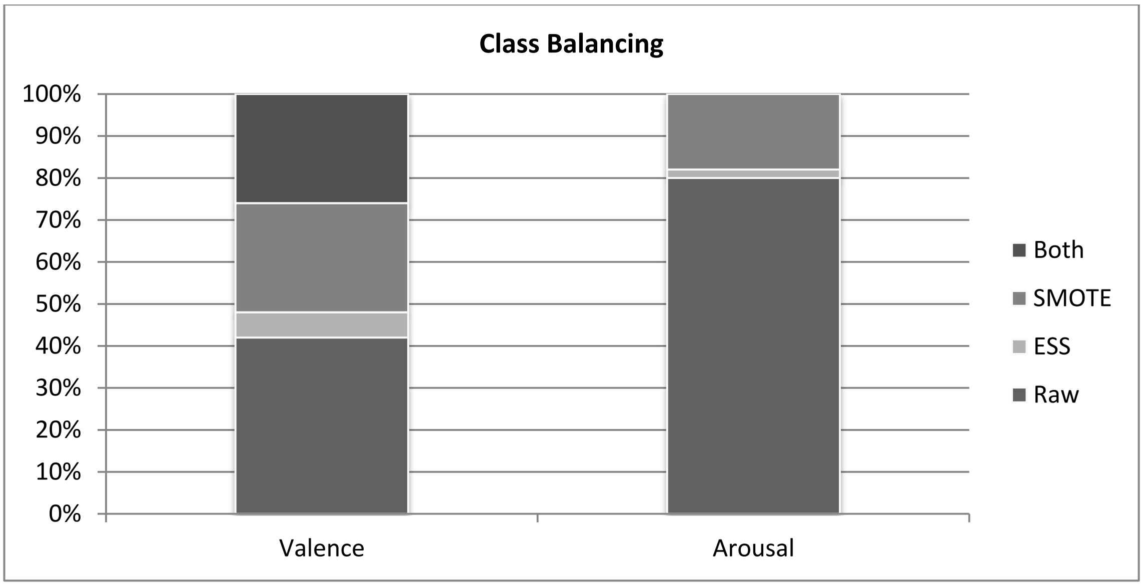


Figure 67. Class balancing techniques used in top 50 models for each emotional dimension

Class balancing techniques, depicted in Figure 67, are much more used in the top models predicting valence (58% of them at least use one class balancing technique) than the models generated for predicting arousal (where only a 20% takes advantage of any of these techniques). SMOTE technique is being used in more top models than ESS.

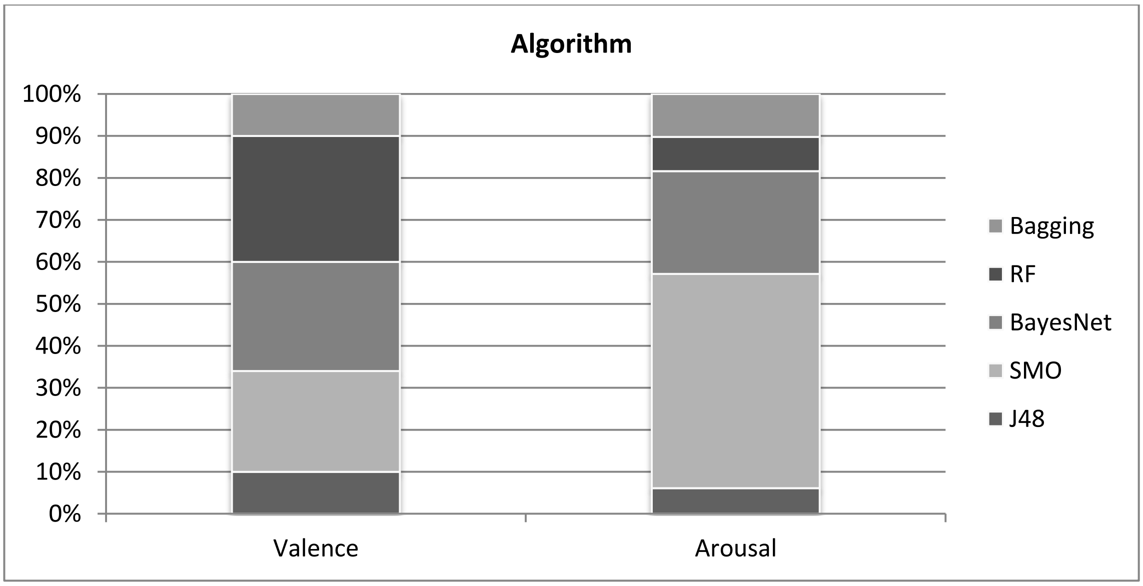


Figure 68.Algorithms used in top 50 models for each emotional dimension

Regarding the algorithms (Figure 68), SMO algorithm and BayesNet are the most frequent algorithms used in the top arousal models, while in the valence models, SMO, BayesNet and RandomForest are the most common algorithms.

In this stage we have also aimed to deal with some of the research questions introduced in section 1.3:

- **Q3**: *Which affective state labeling strategies are more effective in real-world educational scenarios without penalizing aspects such as the intrusiveness of the approach proposed?*
    - Although we addressed this question in stage 1, it has also been partially addressed during this stage 2. In this stage we have evaluated different labeling approaches, regarding different discretization methods on numeric affective labels (see results in section 6.9.4). In this sense we have seen that in the not normalized dataset, using a narrow range for the neutral affective states provided better results. In the normalized dataset, 2-class approach provides a better top model predicting valence, while both proposed approaches provided similar results when predicting arousal.
- **Q4**: *Can the use of reference interaction patterns (collected in a non-affective task) reflecting each individual personal interaction behavior help to improve the affective state detection in real-world learning scenarios?*
    - To address this research question, an initial interaction baseline was included in the design of the experiment. Results (in section 6.9.1) show that using a reference baseline (dynamic baseline) provided better results than not using it when predicting valence. When predicting arousal, the reference baseline did not provide better results.
- **Q5**: *When using reference interaction patterns, can the regular update of those reference patterns help to improve the affective state detection in contrast to using a reference interaction pattern at the beginning of the interaction?*
    - One of the analysis performed in this stage was comparing using a fixed baseline and a reference baseline for data normalization. Only the use of a dynamic baseline, updated after every single task, provided better results than not using any baseline (or using a fixed one) when predicting valence.
- **Q6**: *To what extend the way the multimodal data collected in a real-world learning scenario is handled prior to the model generation with affective state detection purposes can have an impact on the prediction results obtained from that model?*
    - In the results shown in sections 6.9.5 and 6.9.6 we have evaluated the impact of different preprocessing techniques in the model generation stage. Additionally, the interaction data normalization as well as the class- attribute discretization used also have an impact on the models generated. We can see how, depending on the dataset used (how it is normalized and how the labels have been discretized) the different preprocessing techniques can provide slightly better results.

6.10. Discussion on Stage 2 results

As done in stage 1, after finishing the experimentation of this research stage, some of the issues analyzed in this stage are to be discussed:

6.10.1. Interaction baseline model

As previously stated (in section 1.4), one the main hypothesis we wanted to test is that "In real-world learning scenario based inter-subject experiments, the use of a reference state to normalize each user interaction related data provides more robust models when detecting affective features to enrich learner modeling in educational contexts". We created an interaction baseline model that can be used as a reference for how the participant generally interacts with the keyboard and mouse. This way we provide new modeling alternatives which are based on leveraging a user's specific changes and comparing their values across different tasks with respect to the calibration task (i.e., where the baseline model is obtained).

> The initial interaction baseline model's goal is to model keystroke-level features in relation to affect, thus enriching predictive models. This approach follows what is commonly applied when processing physiological signals [189]. Using baseline measurements as a reference model is not uncommon [65,120], but our proposal differs in various ways. We have used a single baseline model to compare the user's behavior over several different tasks in a real-world scenario characterized by the shortage of data.
>
> The calibration task to get the initial baseline model takes little time and is done once at the beginning of the experiment, thus fulfilling another requirement of our approach, which is to provide experimentation settings that closely represent natural learning settings and can be applied in real-world learning scenarios. This way we are trying to minimize the usage of a fixed text, not forcing the user to type it several times, as others have done to collect observable features directly from it [76,228]. This has been an initial proposal that can be adjusted to different experimental conditions. For instance, a keyboard baseline model task could be repeated in a long-term experiment to study the validity of the proposed model over time. The fixed text used in that model can change in order to recalculate baseline features and evaluate their usage. Another possibility in a long-term experiment is to replace this baseline model aggregating new features calculated from a very long time window, such as features from the same digraph typed many times during one day in a free-text data collection approach.
>
> Regarding the choice of the text selected for the initial calibration task, related work used a text from a given popular book, with no apparent reason provided [228]. We took a similar text as in [76] because it provides relatively simple sentence structure with no long uncommon words, and each piece of text has roughly the same length [77]. Besides, it is appropriate for essay writing tasks, which are commonly performed by learners of English as Second Language, the target population in our experiment. The complexity of the proposed text has to be moderate in this case, as copying a text that the participant cannot understand may lead to confusion or frustration. All this raises another issue: the choice of

the text may also impact the quality of the initial model. On the one hand, the shorter the fixed text is, the less intrusive it is for the participant (as the time the participant spends in that calibration task is time that is not being used on a “real” task). On the other hand, the longer the text and the more diverse words included, the more digraphs are modeled (so the model generated will be richer) [1]. So for a given population and educational context a balance has to be found between these two factors. The meaning of the text is another point to address, as it can elicit emotions by itself [2] so a neutral content text has to be selected.

In addition to the initial baseline calibration, a dynamic baseline calibration has also been proposed in this work. In contrast to the initial baseline calibration, the tasks proposed are used as calibration tasks (for each task, the prior task values were used as reference values). This idea differs slightly from the traditional physiological baseline, but seems to provide better results than the initial baseline calibration approach (see Figure 47). The problem of this dynamic baseline calibration is the impact the own task might have on it. Long term experiments would be an appropriate next step to further evaluate both approaches.

6.10.2. Experimental environment-related limitations

In this work we are building models for people in a real-world non-intrusive educational setting from different data sources. There is related evidence showing that models which focus on individual person features tend to be more accurate, but in these settings there is lack of large interaction data sets from which to get an accurate model of the learner features, which is a well-known challenge [122]. This issue has driven us to explore additional modeling features based on different types of measurements. The problem here is that we are not dealing with person dependent models recorded in ideal conditions (usually obtained in non-authentic contexts) but with more naturalistic contexts where lower accuracies are obtained [65,92]. Related work has shown that individual models are difficult to build because of difficulties in getting enough samples per user, and usually, those datasets in which predictions do not surpass certain accuracy thresholds are neglected [76,92,122].

Taking into account the relatively small number of instances available in the dataset, the overall results encourage us to continue to do research in this direction. The kappa values obtained are not so high due to the nature of the experimentation data, which is relatively more representative of a real-world scenario than previous fixed-text data. Hence, further research in this area with bigger datasets could help create more robust models. In this sense, a more robust baseline constructed from more data instances could help us reduce the noise in the data.

Bearing this dimensionality challenge in mind, from the modeling viewpoint, we have included features from keyboard, mouse, physiological signals and performance in order to detect affect state changes. Considering the small amount of instances available in our current dataset, the overall results suggest further research in this direction. Note that in classifier design, some papers suggest that there should be 10 times more instances (training samples) per class than the number of features [105], which are challenging given the circumstances of our setting (which are similar to others [121]). A long term version of the proposed experiment would enable us to extend the approach by providing more exercises over time and thus increase the number of instances in our datasets. Furthermore, instead of dealing with groups of students, focusing on personalized modeling, each participant may be involved in an within-subject study, which is expected to increase accuracy results [122]. This within-subject design would provide a more robust baseline and most likely additional performance features for each research subject, thus enabling to model sequences of actions more accurately and create more complex performance-based models [6].

In any case, the proposed approach in this work provides new modeling opportunities, evaluated in real-world scenarios with multiple users, which can be further explored in future experiments where more interactions could be involved. In particular, we plan to conduct a long-term version of the proposed experiment, using a within-subject approach, to study individual features over different real-world problems. This way, with more interaction data from each learner, we expect to provide a more robust baseline model and represent a wider range of the student's performance across features, to model sequences of actions and create more complex performance-based models [6].

6.10.3. Data preparation

Another key subject is the impact of using a particular data preprocessing method. One of the goals of this preprocessing is the high dimensionality of data in this field, which is a relatively common problem identified in literature [31,76,248]. This is usually tackled by means of different preprocessing and dimensionality reduction techniques. There are many dimensionality reduction techniques and their use can have an impact on different aspects of the model generated, from model interpretation (e.g., PCA generated features are calculated by combining the original ones, so it will not be possible to evaluate which set of original features have the most impact on the results) to the performance of the model generation, as some techniques, like forward feature selection, can be time consuming. In case of class balancing techniques, the use of undersampling based techniques is debatable when there are very few instances in the dataset. In this work some of the most common preprocessing methods have been applied as a variable to be considered when generating the models. The purpose here is to evaluate their impact on the models generated, but usually related work uses them without evaluating their appropriateness or impact (with very rare exceptions such as [27]). In our case, evaluating the class balancing techniques used (section 6.9.5), SMOTE oversampling seems to provide better results than the other techniques used, especially when predicting valence. When evaluating the feature selection techniques used (section 6.9.6) Forward Feature Selection seems to provide similar or slightly better results than not performing feature selection. Nevertheless, it should be discussed the potential withdrawals of using some of these techniques (as forward feature selection can take huge amounts of time). Although they have not been addressed in the work reported here, other technical issues related to the different preprocessing methods should be taken into account, such as time consumption. Due to the small size of our dataset, time consumption has not been a problem, but in large datasets, it could make a difference, particularly when the models proposed are used in other real world scenarios [33].

> Another important aspect of the data preparation is providing the proposed approach with the capability of adapting to different experimentation conditions. By using appropriate data preprocessing techniques, the system might be adapted to tasks where one of the proposed data sources is not being used. For instance, this approach can be employed in educational games where only a mouse is needed [190], thus dealing with having a lack of information from other data sources (e.g. keyboard interactions are not needed or the user does not have physiological sensors) [31]. This approach can also be explored to adapt affect detection to people with special needs. In this case, [211] suggests that some of the proposed data sources used in this work may change their interaction purposes. For instance, blind people use a keyboard for navigating over the materials with a screen reader, and we could detect and compare their keystroke behavior when they are either navigating or writing [211].

To tackle this issue, different classifiers could be used for each data source and the different models generated could be combined, thus determining which data source could offer more information for each user in a given situation over time. The use of unsupervised learning techniques can also be integrated with the research described here, thus generating groups of similar users and generating models for the users depending on their group, following related approaches in exploratory learning environments [80].

6.10.4. Discussion summary

As we have seen, new methodological variables have been analyzed in this research stage. We have seen how the interaction baseline model can provide better results in some circumstances, but its further application requires a deeper research in many related issues in order to collect that baseline. Also new scenarios, such as the intra-subject approach could open new elements to analyze in this sense.

Regarding the evaluation of the AMO-ML methodology drawn in this work to a real-world learning scenario, there are still many points to address (to be done in section 6.10.4). The inclusion of this context has driven us to simplify some issues addressed in the experimentation in stage 1, but new issues have raised.

Regarding the data preparation, there is still a lot of work to do in this sense. Due to the immensity of the data mining field, there is a wide variety of preprocessing techniques to be applied, each one to be suitable to different situations. In this stage we have evaluated only some of the most common techniques used in the field, but this research could be easily extended.

7. Summary of Results Obtained

Starting from the AMO-ML methodology developed in stage 1, and including the lessons learnt in the transition stage, new methodological questions arose. In order to further evaluate those new methodological issues, we have designed and carried out several experiments. Thanks to those experiments, the methodological approach proposed in stage 1 has been used and improved to its use in real-world learning scenarios (looking for an ecological validity). This AMO-ML methodology is based in a machine learning system capable to predict the affective states. With this methodology developed during the different stages of this work, we aimed to address the first research question introduced in section 1.3 (**Q0**: *Can machine learning techniques be used in order to detect learner's affective states in realistic learning scenarios from data collected from different data sources?*). As we have seen in the results of stage 2, it is possible to perform affective state detection in realistic learning scenarios from data collected from different data sources. Nevertheless, there is still room for improvement, as there are many new approaches to evaluate from the methodological variables identified. To get there, many approaches identified in literature during an initial exploratory analysis have been followed and some methodological variables found in the field have been evaluated. In stage 1, we have identified the robustness of multimodal approaches in contrast to approaches based on one single signal, aiming this way to confirm the first hypothesis proposed in this work (**H1** in section 1.4). In order to validate that hypothesis, all the possible combinations of data sources have been performed and their results evaluated (objective **O1.1** in section 1.4).We have also evaluated several labeling approaches (objective **O1.2** in section 1.4) involving external annotators with different backgrounds, with the best models generated when performing valence prediction (sections 4.9.1, 4.9.2 and 4.9.6). These best results might have been derived from some methodological issues (e.g. the design of the sentiment analysis) that were addressed in stage 2 and will be discussed in section 8. Nevertheless, the main result of the first stage of this Thesis has been the development of the AMO-ML methodology and the infrastructure of a system capable to perform affective state detection in learners.

Stage 2 has been designed to provide more ecological validity to the approach proposed in stage 1. In order to do that, we have built a similar approach to the one used in stage 1, taking into account the lessons learnt in that stage. The main differences were the inclusion of a real-world learning scenario based on the references set in the transition stage, as well as the use of an interaction data normalization approach. These two differences were introduced in order to validate the second hypothesis defined in this work (**H2** in section 1.4). Nevertheless, some changes were introduced in stage 2 in order to move the approach to a real-world scenario. In addition, in stage 2, other methodological variables have been evaluated (depicted in Table 27). From this second stage, the following conclusions can be drawn from the results reported above: i) the use of a baseline in order to calculate mouse and keyboard normalized features may offer better accuracy rates when predicting affective valence (to address the first objective **O2.1** defined for the stage 2 in section 1.4); ii) when using discretization method with the class attribute (objective **O2.3** defined for the stage 2 in section 1.4), those approaches with an unbalanced bin (e.g. a small neutral bin with positive and negative bins representing a wider range from the original variable spectrum) tend to offer worse accuracy rates and iii) the only dimensionality reduction technique that provided better results was forward feature selection (while SMOTE oversampling, equal size sampling and PCA offered worse results than not reducing the dimensionality). These last two points were evaluated in order to face the second objective (**O2.2**) defined for stage 2 in section 1.4.

To get these results, a wide variety of issues have been researched. Since our first approach developed in stage 1, a lot of methodological variables that may have an impact on the prediction results have been evaluated. Table 28 shows the different methodological variables evaluated identified (within their corresponding methodological aspect), together with the approaches chosen to address that variable in each different stage of this work. In addition, some final remarks about the evaluation of those methodological variables are included focusing on the results obtained in the stage where that variable has been evaluated.

Methodological aspect	**Methodological variable**	**Stage 1**	**Stage 2**	**Transition Stage**	**Remarks**
Characterizing and labeling affective state	labeler	Participant, educational expert, psychologist	Participant	Psychologist	In stage 1, the best results were provided by a mixed approach based on the combination of the SAM scores given by the participants with the labeling provided by the psychologists. Results: see section 4.9.8.
	Time of the labeling	During the experiment (participant) and after the experiment (experts)	During the experiment	After the experiment	In stage 1, the best results were provided by mixing the labels collected during the experiment (SAM scores provided by the participant) and after the experiment (psychologists labeling). Results: see section 4.9.8.

Methodological aspect	Methodological variable	Stage 1	Stage 2	Transition Stage	Remarks
	labeling approach	SAM (participant), plain text (participant), scores (experts)	SAM	Categorical labels	Although in the ITS experiment a categorical approach was followed, most of the research here presented has made use of a dimensional one, using the valence and arousal scores collected by means of the SAM scores.
	affective state characterization	Positive-negative	Positive/neutral/negative (if no data normalization is applied); Positive/neutral/negative (if data normalization is applied); not-negative/negative (if data normalization is applied); neutral/not-neutral (if 2-step classification is used).	Different categorical emotions (anxiety, confused, concentrated, frustrated, happy, shame or surprise, none)	In stage 2 results, we could see how, in case the data was normalized, the better results were obtained when predicting “not negative/negative”. If the data was not normalized, the best prediction results were obtained when the neutral category containing only those instances scored as 5 out of 9 in the SAM scale. Results: see section 6.9.4.
	Time window labeled	Group of tasks	Single tasks	Fixed time (after evaluating different fixed time windows)	Due to the educational context of this research and the emotional elicitators used,in stages 1 and 2 we used a task-related time window. The fixed time window approach used in the transition stage is more appropriate for task independent approaches.
Data processing	normalization	Using personalized baseline (only physiological signals)	Using personalized fixed baseline (all data sources and labeling); using personalized dynamic baseline (all data sources and labeling)	Using personalized baseline (only physiological signals)	When predicting valence, a dynamic baseline based approach provides best results, while when predicting arousal, the not normalized dataset provided the best models. Results: see section 6.9.1. When choosing the baseline to be used with the physiological signals, the combination of the initial and final baseline recorded provides the best results in both valence and arousal. Results: see section 6.9.8.

Methodological aspect	Methodological variable	Stage 1	Stage 2	Transition Stage	Remarks
	preprocessing techniques	Correlation Filter Low variance filter	Correlation Filter, Forward Feature Selection (FFS), Principal Component Analysis, SMOTE, Equal Size Sampling Low variance filter	Backward Feature Elimination (BFE), Principal component Analysis (PCA)	Regarding the class balancing techniques evaluated, SMOTE seems to provide the datasets with best results. Results: see section 6.9.5. Regarding feature selection technique, it seems that forward-feature selection technique provides best results in most cases. Results: see section 6.9.6.
	Algorithm	J48 Naïve Bayes Random Forest SVM Bagging Bayesian Network Neural network	J48 Naïve Bayes Random Forests SMO Bagging Bayesian Network	J48 Naïve Bayes Bagging	In stage 1, J48 and random forests were the algorithms that provided more top models. Results: see section 4.9.9. In stage 2, Bagging, Bayes net and SMO seem to provide the best results (but not significantly better than the other algorithms). Results: see section 6.9.7.
	Stepwise prediction		2-step classification, Clustering	2-step classification	Although in the ITS experiment the use of the 2-step classification approach seemed promising, in stage 2 the results do not provide better results when performing it. Results: see section 6.9.3. Regarding the use of clustering techniques prior to the supervised learning model generation, it seems that not using any of the proposed clustering techniques offers better or similar results than using them. Results: see section 6.9.2.
Experimental approach	context	Laboratory conditions	Real classroom & laboratory conditions	Real classroom	This variable has not been evaluated.
	task proposed	Math problems	Essay writing in English as a Second Language	Math problems	This variable has not been evaluated.
	elicitation methods	Time limit, difficulty	Time limit, difficulty	Time limit, difficulty	This variable has not been evaluated.

Methodological aspect	Methodological variable	Stage 1	Stage 2	Transition Stage	Remarks
	data sources	Keyboard, mouse, physiological signals (heart rate, GSR, skin temperature, breathing), sentiment analysis	Keyboard, mouse, physiological signals (heart rate, GSR, skin temperature, breathing), sentiment analysis	Keyboard, mouse, physiological signals (heart rate, GSR, skin temperature, breathing)	This variable has not been evaluated.
	educational task	dotLRN test	MOKEETO	ITS developed by University of Valencia	This variable has not been evaluated.
	Participants	78	41	2	This variable has not been evaluated.

Table 28. Methodological issues addressed in this work and the different approaches followed in the stages of this research

As we can see in Table 28, one of the main contributions of the AMO-ML methodology here proposed in contrast with related works is the comparison between different possibilities in the different methodological issues evaluated. This provides a base for the methodological decisions taken as a deep description on all the issues that have raised whit every possibility evaluated. That description aims to be useful in the design stage of further related approaches, in order to help to take informed decisions on many different aspects of the development of those approaches.

Regarding the impact of the work here proposed might have on the learner, the work here described opens (as well as all the related works in affective state detection) a wide range of opportunities to improve the learning experience. The idea behind providing an affective state model of the learner is to allow other systems to take advantage of that information. Those systems that take advantage of that emotional model of the learner can provide a better and more tailored learning experience (e.g. recommender systems based with affective capabilities or systems that might teach emotional self-control). Although the main outcomes of this thesis might not impact directly on the learner (as this work requires a system to use the emotional model generated), some issues taken into account in the AMO-ML methodology here proposed have been designed to avoid a negative impact on the learner (e.g. the intrusiveness).

8. Discussion

We have identified and further explored many modeling issues involved in unobtrusively detecting the affective state of a learner following a multimodal approach. During the proposed experiments, we have automatically created multiple models using machine learning techniques where we have combined features and data sources proposed in different research works and extended them in order to address a real-world learning scenario.

As we have seen along the previous sections, there are a lot of facts to take into account when designing experiments for detecting emotions in an educational scenario, and all of them may affect the results. As a result of the research work carried out and presented here, there are many things to discuss, especially after the experiment results have been analyzed, when we can see the consequences of all the decisions taken.

8.1. Affective state representation

Another variable addressed in the current work is which affective model representation to use [64]. A simplified dimensional approach for labeling the user's affective state has been used, which is readily available and can be managed by students themselves [39]. The problem of labeling affect is well known, and here inter-observer agreement can be low [64]. The use of sets of emotions, where the participant has to choose the closest to her current state, is common in the literature (as seen in section 2.3.1.a and as used in section 5.2.7). This categorical approach can result in different interpretations of the same emotion [120]. We have dealt with the potential problem of affective concept misinterpretation by providing learners with a written explanation of the two affect dimensions covered in this approach, valence and arousal, in order to ensure that all learners understand the dimensions in the same way. Other approaches deal with this issue by providing mappings from affective dimensions to affective categories, even when they are focused on learning-centered cognitive-affective states [20]. Although this categorical approach is out of our main scope for the reasons discussed above, there is evidence that coders who have received proper thorough methodological training assessing affect have achieved inter-rate reliability rates of over 0.6 [162].

> Our work has explored different ways of discretizing the dimensional values obtained into different dimensional categories to perform the prediction. This approach seems to be interesting since it provides a finer-grained means of modeling students' affective states. However, this would hamper the simplicity we are aiming at. We have selected a discretized approach that simplifies the

problem of predicting the dimensional values of the affective state for the algorithms used. By using positive, negative or neutral states (from the raw dataset) or positive or negative emotional transitions (from the user-normalized dataset) we are keeping a reduced number of categories, which are easier to manage when there is a shortage of data. In this sense, more discretization approaches could have been evaluated, thus following this paper’s approach, which shows the first step at looking for a proposal that supports classification algorithms and also is able to provide a balance between meaningfulness and simplicity for users.

The emotional modeling approach deserves particular attention when the affective state detector is to be integrated with other components that will use the predictions performed [13]. This is another reason why this dimensional approach was chosen, as it provides a more flexible and standardized description to select which affective state phenomena to take into account. Although some previous works use high/low or positive/negative values for the proposed dimensions [79,116,254,255], the use of neutral states could also be taken into account, for example, to use the two-step classification approach used to discard those neutral states before classifying positive or negative affective states [192].

Regarding the evaluation of the work done in the discretization approaches for the labeling attribute, further research could be performed, in which a recommender system employs the different approaches evaluated. A recommender system would give us a closer look at the impact of discretization criteria on the recommendations given, towards understanding what granularity is most important in different situations. The characterization of affect determines the type of recommendations that can be implemented in real-world scenarios. Recommendations could be provided in terms of well-known traditional interaction sources, such as the ones used here (i.e., mouse and keyboard) or less explored interaction sources, such as visual, sound or haptic, which can be tuned to provide recommendations in ambient intelligent scenarios [71,215].

8.2. Data sources

One of the variables evaluated on the first stage was the different configuration approaches to be generated from the data sources proposed. Although the most common data sources in related works are the physiological signals, due to the educational context where this work is being carried out, the inclusion of keyboard and mouse was considered appropriate. A key issue in this work is the suitability of these interaction devices that are used and modeled. During the stage 1 of this research there were very few keyboard interactions during the tasks, so the following experiments were oriented towards more keyboard-related tasks.

Although the task in stage 2 mainly depends on typing, the mouse was used by all the participants for navigation and for text editing purposes. In a series of tasks where the keyboard is rarely used, automatic feature selection methods should discard the keyboard features. Similarly, a judicious selection of threshold values (like the ones depicted in Figure 40), also ensures that features are only used when they represent an interaction repeated a potentially significant number of times.

> As to the model generation and the interactions device used, although all the participants interacted through keyboard and mouse in this work, it is common nowadays to find users controlling desktop computers (or other devices) by other means, such as track pads or touchscreens. The link between touchscreen based typing and mood changes was investigated in [175], which uses digraphs and trigraphs and takes into account other touchscreen related features, such as the number of hands used for typing. In [21], the use of a touchscreen is also combined with keyboard and mouse features for affective computing.

Regarding the physiological devices used, as aforementioned, that is one of the most common approaches followed nowadays in affective computing works [205]. During the different stages of this work different devices have been used. Starting from a commercial device for physiological signals recording, which offered accurate values, but with some drawbacks, such as the price availability, the intrusiveness or the impossibility to manage the data recorded live. Many works use similar devices, more research oriented, expensive and similarly intrusive [125,184], but in stage 2, one of our goals was take the experimental approach followed in stage 1 to a real-world educational context. Although the keyboard and mouse were common in real-world educational contexts, the physiological signals were not. Some efforts were carried out to make the physiological sensors less intrusive and more usable. With the development of an open hardware-based solution, the cost was reduced notably, and some sensors were changed to make them less intrusive (heart rate sensor was changed, as in stage 1 participants had to wear electrocardiography sensors, needing to wear stickers attached to their body and in stage 2 heart rate was obtained by means of a photoplethysmography sensor attached to the ear lobe with no need to stick anything to the participant's body). Although the initial idea of this work was using current wearable devices capable of collecting most of the signals proposed, due to the new development of the AICARP platform in the research group where this work is framed, it was decided to adopt it in the stage 2 experiments. This adoption also meant losing some physiological data due to the early stage of the device in the first experiment of stage 2 (as described in section 6.7.2). Further experiments should be carried out with wearable devices, allowing participants to provide the physiological data proposed in this work only wearing a smart bracelet.

8.3. Time window

As this work is also framed in the educational field, it is hard to create that emotional impact on the participants (commonly used stimuli include movies, images or other non-educational-related situations). The commonly faced scenario in emotional experiments in educational contexts are usually designed to detect frustration or boredom, emotions that arise in long term tasks, which makes it harder not only to isolate the moment of the emotion but also to guarantee that that emotion is 1) caused exclusively by the proposed task, and 2) is just the emotion expected and not a mix of different emotions (as in other fields, the emotions elicited are elicitated as an immediate reaction to an impulse). In summary, creating “isolated” emotions to detect in an educational context is something quite difficult where many steps should be taken.

In stage 1, instead of taking a timestamp every time a new problem was shown, the timestamps were only taken at the beginning and at the end of each group of problems, being this, a huge time window where many different emotions could have appeared. This may conduct to time windows were many emotions could be reflected, but only one class value (which may represent the last felt emotion, the strongest emotion of the task or whatever). That is why finally we decided to use only the rows representing the emotional report tasks (at the end of each group of problems), as during the time the participant express his or her emotions, there is nothing that disturb the participant that may change the expressed emotions (but the mere act of expressing them, but that is something that will be always present when asking participants for feedback about their emotions), so it may be the closest we are to a “isolated” emotion. Anyway, the sentiment analysis score that has provided so much affective information when detecting emotions (as seen in previous section), should be calculated from task-related texts instead of making learners stop to type their emotions. This should be done with a reconsideration of the approach adopted as in this work a first attempt was carried out.

This timestamp issue (i.e., taking the experiment timestamps from a device which does not use the same time reference than the other devices used) could have been solved automatically with a device that allowed to take automatic timestamps triggered by a signal (maybe a signal generated by the server every time a problem page is loaded), so that is important to know well the available devices to use and, in case the devices are going to be bought or developed, that functionality should be present to avoid human errors on timestamp taking.

In stage 2, this issue was faced taking timestamps every new task. To do that, the tool developed to perform the tasks (MOKEETO) generated a log with all the task times according to the local pc time. This issue has also to be taken into account when designing the tasks proposed, as the task duration should be long enough to manifest the elicited emotion avoiding other possible emotions that could arise in excessively long tasks (e.g. boredom, frustration, etc. in case those are not the elicited emotions).

In the experiment described in section 5, a different approach was followed. As that experiment was carried out in collaboration with people from University of Valencia, their ITS was used. Regarding the data analysis carried out, two collaborators with the aDeNu research group carried out an ANOVA was conducted for each of the temporal windows of the problem and the last temporal window included in the final baseline, indicating which temporal windows (per subject and signal) were significantly different from the final baseline ($p < 0.001$). This lead us to handle a fixed time temporal window approach, based on the statistical analysis performed to the physiological signals. This was used as in this experiment, the labeling used was carried out by an external annotator who watched all the experiment videos and had no temporal constraints when labeling, adapting the annotations performed by the labeler to the temporal windows defined by the ANOVA. This approach was not used in the stage 2 in order not to require of any external annotator to label the affective states of the participants.

8.4.Live data processing and data synchronization

Another desired functionality is the streaming and analysis of the data live. In this work, all the data analysis has been performed after the experiment. In stage 1, this approach was chosen due to the limitations provided by the physiological signal recording device used. Although the recording device was different in stage 2, the device had its own recording program. The live processing of the data would arise new methodological issues as well as infrastructure complications such as if the data processing is to be performed in the participant's computer (with the possible computing limitations of that computer) or in a data analysis server where all the data is sent live (taking into account issues such as the privacy or the volume of the data sent). This issue is related to the aforementioned issue of the data synchronization. In stage 1 (using the dotLRN platform), a bad management of the synchronization of the data limited the time window to evaluate to the emotional report. In stage 2, the tool implemented for the task proposed, generated an automatic log with the timestamps of all the actions performed in the tool, including task beginnings and ends. In case the data analysis server approach is to be carried out, context information (in this case, from the task being performed) should also be sent in order to be used.

8.5.Elicitation methods

It was seen in section 2.4 the wide variety of elicitation methods in related works. Due to the nature of the context of this research (the educational context), some of these methods did not fit the scope of the research here presented. Nevertheless, there are some factors that should be discussed about the elicitation methods used in the different experiments here described.

Regarding the data collection, the impact of the task proposed on the data collection is an important factor to take into account. In stage 1, it should be discussed the dependency on the mouse and keyboard behaviors to the task proposed. On the one hand, during emotional reports tasks, an intense use of the keyboard was needed, while on the other hand, in the problems tasks, using the mouse could almost be enough to solve all the problems (although keyboard was needed to provide the SAM scores). Something similar happened in the experiment described in section 5, where the ITS developed by University of Valencia was used. Initially, the ITS barely needed the use of keyboard in order to solve the proposed problems, but people from University of Valencia introduced some changes to make mandatory the use of the keyboard (by means of explaining each variable declared). That led the design of the task proposed in stage 2. An educational-related task where the use of the keyboard was needed. With the inclusion of this task, the use of the mouse was also needed with navigation purposes as well as with text edition purposes.

Another important point related to the elicitation method, connected to the problem described in the previous paragraph, is not only the task proposed, but also the way the elicitation materials are presented. In stage 1, two clear presentation errors were committed. First, when presenting the images from the IAPS, the images were presented in a dotLRN questionnaire, in a web page. Usually, these images are shown in a room, with certain light conditions presented in full screen for a concrete given time. However, in our experiment, while the image was being shown, participants could also see the dotLRN interface, reducing the emotional impact of the images. Participants were also told to skip the images after 5 seconds of viewing, but they usually skipped it after 2-3 seconds (specially the hard ones), which can reduce the influence of the image on the participant. As the images to elicit emotions were no longer used, this issue was not taken into account anymore. The second presentation error was during task 4 of the stage 1, where there was a time limit for solving all the problems. In order to boost the stress sensation, the countdown should have been clearly shown to the participant, but only a non-dynamic countdown (which values were only updated when a page was loaded) was used during the experience as it was implemented in dotLRN that way. This issue was taken into account when designing the countdown in the MOKEETO application, so, in the experiments held in stage 2, the countdown was shown in a dynamic way, including milliseconds to induce more stress on the participants.

Regarding the materials used in our experiments in order to elicit emotions, it should be discussed the suitability of the proposed materials. Due to the wide variety of participants in stage 1, it was difficult to design an experiment personalized to the current mathematical skills of each participant. In that stage, when some groups of participants coming from a school came to participate, the materials were adapted. As the experiment described in section 5 was designed to be held in a real classroom, the materials were carefully chosen (by two collaborators from the aDeNu Research Group) and the difficulty of each problem was evaluated in order to design a proper flow of problems in order to elicit the emotions. Something similar was done in the experiment held in stage 2, as the materials for the experiment (the proposed vocabulary) was extracted from the textbook the participants from the school were using for their regular classes in order to adapt the difficulty to the participants.

Another point that could be discussed is if the elicitation methods work the same way for all the participants. Data mining looks for patterns in big datasets containing the key attributes that generate the value to be predicted, but in affective computing, there are many works as have been seen in the state of the art aiming in many different directions, but most of them aiming to an inter-subject approach, assuming that there is a common pattern in affective behavior. To continue in this way, a huge amount of data is needed, from the most heterogeneous sample possible, and as we have seen in this work, it is not easy to collect this data. Nowadays, thanks to MOOCs, is “easy” to generate a course with thousands of learners, but the collection of data from them (especially the data proposed here) is still an intrusive issue (both at the physical and privacy levels).

Another plausible approach is to adopt an intra-subject point of view. As mentioned in previous paragraph, most works nowadays rely on an inter-subject approach, but some of them are starting to focus on the modeling one or very few subjects [63,130] in a long time period, which may help to develop an personalized affective model. An evaluation of an experimental design over both (inter-subject and intra-subject) approaches would be interesting take in order to evaluate the most proper way to carry out affective computing experiments. This has been one of the issues that lead us to evaluate the different normalization approaches used in stage 2, in order to try to normalize each user data using its own baseline, trying to get closer to an intra-subject approach from an inter-subject experiment.

8.6. Accesibility

From the accessibility point of view, it has to be mentioned the importance that issue should have in this research. During stage 1 an experimental session was carried out with two participants with visual impairment (one partially blind subject and other totally blind). The adaptations were reported in [211]. Regarding the processing required in that kind of scenarios, a personalized way of processing some data sources information depending on some possible special behaviors shown by people with some impairments or having certain medications should be considered. The intake of some medicines may affect some physiological signals, and blind people may show some movements that may affect some of the measures taken. This issue makes arise again the possibility of moving forward in the direction of an intra-subject approach, where a detailed model of each person is performed. Two possible ways of acting arise from this point, one of them is using a special way of processing some data sources that may be influenced by some special behaviors related to a physical or psychological condition or the other is, as this work proposes, rely on a strong multimodal approach, letting the system itself drop those data sources that, due to a special characteristic, introduce noisy data as they are “not being used as expected”, being used only those data sources that may not be influenced by that special characteristic.

8.7. Privacy

Another point to take into account, due to the nature of the data handled in this field, is the privacy of the data. In this sense, some concerns should be evaluated. First, the user should be aware of what affective aspects are being modeled, and how those aspects of the user model can be used. As we have said before, emotions play a key role in learning, but it also plays a key role in other aspects, such as marketing or even politics [161]. In case some of the work here explained is used in any platform, the user should be aware of the use the data collected is going to have and how it is going to be processed. Another point to face regarding to the privacy, is, when AMO-ML methodology is applied in a deployed system, when and where the data collected is going to be processed. In case the data is going to be generated in an external server, the keyboard interaction data, for instance, should be sent to that server, sending, this way, sensitive data such as passwords or other kind of information. In contrast to that, processing the data in the participant device (in order not to send possible sensitive data) might derive in a high computational consumption in the client level. This kind of scenarios forces to take into account the importance of the data privacy and how to handle the processing of the data collected (evaluating the sensitiveness of the data).

Regarding this issue in the research context where this work is framed, a series of concerns should be minded too. When carrying out research experiments, the issues described in this section should also be taken into account. Many institutions have a unit responsible for bioethics related issues. In the case of the experiments here described, they had to be approved by the bioethics committee from the institution where this research is being carried out (UNED). After the approval, every subject had to sign an informed consent form or get it signed by their legal representative in case the subjects are minors (See section 13.2.1).

9. Contributions

The main goal of this work has been the definition of a methodological approach to perform affective state detection using machine learning techniques. This AMO-ML methodology has been built from an initial exploratory analysis (and described in an instantiation performed in the experiment carried out in stage 1) and has been refined until its application in a real-world learning scenario (in stage 2).

Additionally, during the different stages of the development of the proposed research, many open methodological issues have raised. The analysis of some of those methodological points identified all along the different stages aim also to be a contribution to the field of affective computing. Some tools have also been developed, and predictive models have been generated. All these works have been depicted in the current document.

In summary, the main contributions from this work include:

- An affective computing experimentation methodology (AMO-ML): This Ph.D. thesis has described the process of design, improvement and evaluation of a methodology to be followed in order to perform affective states detection in real-world based educational scenarios by means of machine learning techniques. An multimodal point of view was proposed aiming to evaluate the added value of that approach, addressing this way the first hypothesis proposed in this work (**H1** in section 1.4). The development of this methodology has been carried out in in the two main stages proposed in the current work, following an incremental approach (generating an initial version in stage 1, as a result of an initial exploratory analysis and an improved version in stage 2, defining new methodological issues found in the previous experiments and being applied in a real world-based learning scenario experiment).
- An evaluation over certain methodological aspects and their impact con affective computing scenarios: During the different experiments carried out in the definition, improvement and evaluation of the AMO-ML methodology described in the previous bullet, many methodological open issues have raised. In order to provide a clear view of the impact that those methodological points might have on the results obtained, an evaluation has been carried out on each methodological variable proposed. The experimentations carried out in this work have used different approaches in the following aspects.
    - Data sources used: Many different data sources were proposed and an evaluation on the different data sources possible combinations has been carried out in stag 1 (objective **O1.1** in section 1.4).

- Labeling related issues: Labeling data with affective purposes in order to train supervised learning techniques is a required task in this kind of experiments. Nevertheless, the labeling requires a series of methodological definitions in order to be carried out. During stage 1, the research focus was set in this point (objective **O1.2** in section 1.4). These definitions include:
    - Labeler: The subject to provide the affective labels to be predicted by the system. In this work we have carried out a comparison between the results obtained from labeling approaches performed by different sources (external experts and subjects themselves).
    - Time of the labeling: During the different experiments, different time windows have been evaluated in order to label the affective state of the participants. While in stage 1, the labels corresponded to a series of mathematical problems, in stage 2, each label corresponded to a single task (essay writing). In the transition stage, a fixed time windows was used.
    - Label format: In both stage 1 and stage 2 a dimensional approach was used in order to represent the affective state of the participants. In the transition stage a categorical approach was followed in contrast to the other two stages. Additionally, different discretization approaches over the labeling data were evaluated in stage 2 (objective **O2.3** in section 1.4)
- Experimental context: While the experiment held in stage 1 was performed in lab-conditions, the experiment held in stage 2 was also carried out in a real-world learning scenario. That change of context was initially performed in the transition stage, which helped to define a reference scenario to translate the initial version of the methodology to a real-world learning scenario in the stage 2.
- Emotion elicitation method: Regarding the emotion elicitation method, time and difficulty have been the main resources used in order to elicit emotions during the experiments proposed. This is due to the limitations the educational context provides.
- Tasks proposed: Framed in the educational contexts, different tasks have been proposed: while in stage 1 a series of math problems were proposed (and in the transition stage, but carried out in a different tool), in stage 2 participants were asked to write essays in an English as a Second Language context.
- Different data preprocessing techniques: Data preprocessing is one of the most important steps to take when using machine learning techniques. While some techniques might be used in related works, the impact these preprocessing steps are rarely evaluated (objective **O2.2** in section 1.4). In this sense, some of the most common preprocessing techniques have been evaluated:
    - On class balancing: SMOTE oversampling and Equal Size Sampling have been used and evaluated in stage 2

    - On dimensionality reduction: Forward Feature Selection and Principal Component Analysis have been used and evaluated in stage 2.
    - Data normalization: Different approaches in data normalization have been evaluated in stage 2 (objective **O2.1** in section 1.4), evaluating, for instance, issues such as the reference values to use when normalizing data (comparing using a fixed value and a dynamic value as baseline).
  - Different machine learning algorithms: As done in many related works, different machine learning algorithms were used and compared when generating the affective state predictions.
- The application of an initial baseline (commonly carried out in experiments with physiological signals) to data collected via interaction devices (i.e. mouse and keyboard):

  By mean of this, this work aims to get rid of some subject related bias such as the participant’s keyboard and mouse interaction skills when using data from several participants in an inter-subject experiment, in order to evaluate the second hypothesis proposed in this work (**H2** in section 1.4).
- A series of tools have been developed:
  - A key logger and mouse tracker in order to collect, in a transparent way, data from the interactions carried out by computer users.
  - A data synchronization tool in order to help the multimodal data collection when several devices and computers are needed.
  - MOKEETO: A tool for essay writing to be used in educational scenarios, allowing proposed words to be memorized or not and countdown.
  - A tool for feature generation from mouse and keyboard interactions and physiological signals.
  - An ad-hoc designed data analysis workflow for each one of the stages of this work where the model generation was automated according to the variables evaluated in each stage.

Some of the tools developed in this Ph.D. Thesis are expected to be improved and shared by the author in a repository. A CD with the data analysis workflow and some of the tools here developed are to be provided with this Thesis. Any work produced by partially or the total part of the developments provided by Sergio will indicate the authorship of the materials used.

Additionally, some parts of the work described in this thesis have been published in different journal papers, conferences, etc. A full list of the different works published related to this research can be found in Appendix I (section 13.1).

10. Conclusions

The problem addressed in this work is to evaluate how different methodological aspects may impact on the performance of automated affective state detection systems in educational environments by combining information gathered from several input sources using supervised learning techniques.

In order to carry out this work, two hypothesis were proposed as well as five different research objectives (reported in section 1.4), following a 2-stage experimental-based methodological approach that goes along the whole research cycle. This research cycle has been reported in the previous sections, as follows: the review of the state of the art was done in Section 2 in order to have a clear view of the field of emotion detection in computing, detecting open issues, successful approaches and research lines to follow. In particular, identifying which data sources to be used with affective purposes, with special emphasis on those appropriate for educational scenarios and techniques commonly used to extract information. Selected data sources were keyboard and mouse interactions, physiological signals obtained from bio-feedback devices and sentiment analysis, and the data analysis techniques to be used were supervised learning techniques.

With that information, a work plan to follow was proposed. This work plan, described in Section 3, defines the different experimental steps to take in order to evaluate the hypotheses and objectives proposed. As this work has been designed as a 2-stage research, the proposed steps have been proposed (in section 1.5) to be followed in each one of the 2-stages, having in each one of the stages different objectives.

After that, an experiment was carried out in order to get the data generated from the selected data sources following a multimodal approach, during an educational experience in order to achieve the goals proposed for stage 1. 78 participants participated in the study. The experiment was designed to elicit emotions in a context with an educational charge, as consisted on a series of mathematical tasks, and was designed to collect data from many of the data sources identified in the field.

The data gathered was processed and 735 different analysis were done on combinations regarding the labeling approach, the data sources considered and the data mining techniques, evaluating the approach followed to extract affective information with a ranking score that considers the accuracy and the Cohen's kappa.

As a result, after carrying out the experiment and analyzing the data, we have seen that in 17 out of the 21 top predictions performed the data used as input came from a combination of different data sources. This suggests that the combination of data sources offers better or similar accuracy rates than a single source approach when one device performs better than the combination. This is an indicator that coincides with the first hypothesis stated in section 1.4 (**H1**). To get here, the most commonly used data sources detected in literature have been used (**O1.1** in section 1.4), having in mind price and intrusiveness (although the physiological sensors used during the experiment were a little intrusive, the measures used in this work could nowadays be recorded by means of non-intrusive devices such as strap). Another issue evaluated in that first stage has been the source of the labeling used (**O1.2** in section 1.4). Three different emotional labeling sources (and an approach based on the combination of two of them) have been used in the experiment proposed in stage 1: i) one provided by the participant, ii) another coming from two psychologists and iii) another coming from an e-Learning expert. In this sense, the combination of the labeling provided by the participant with the labeling provided by the psychologists seemed to provide the best results both in predicting valence and arousal (see section 4.9.8).

After the end of the stage 1, and using the lessons learnt from that experiment (as seen in section 6.2.1), the stage 2 experiment was designed. This stage 2 also aimed to take the approach proposed in stage 1 closer to a real-life educational scenario, validating it in a real classroom. The proposed research methodology in section 1.5 was also followed in this second experimental iteration. In this stage, the focus was set to more fine grained methodological variables, especially related to the data processing. A second hypothesis was proposed for this stage, this time aiming to improve the affective state detection rates by mean of evaluating new approaches for interaction data normalization using a personalized baseline (**H2** in section 1.4). Results pointed that the use of a dynamic interaction baseline (i.e. using the last interactions performed by the participants as reference to evaluate her current interactions) provides better affective state prediction results when predicting the affective valence, while not using any baseline seemed to provide the best results when predicting the affective arousal (see section 6.10.1). This baseline approach was inspired by the data normalization commonly performed when using physiological signals [173], aiming to get rid of the impact some interaction related issues might have on models generated following an inter-subject approach, such as the variability of the user skill when performing the interaction with the proposed devices. The different ways to evaluate that normalization have been taken into account in the experimental design (**O2.1** in section 1.4). Also, other fine grained data analysis methodological issues identified in the review of related works have been taken into account. Preprocessing techniques are used by some related works, and the impact of their use on the results is rarely taken into account (as seen in section 2.2). That is why different class balancing and feature selection techniques have been used and their results compared in order to evaluate their suitability in the field of this research (**O2.2** in section 1.4). In this sense, SMOTE class balancing seemed to provide best results on the class balancing techniques. Regarding the data labeling, in contrast with stage 1, where the focus was set on the source of the labeling, the focus has been set to evaluate how different ways of discretizing the affective labels used (**O2.3** in section 1.4).

11. Future works

The work here presented aims to be another step in improving emotions detection in educational scenarios. However there is still work to be done. Part of it is being addressed within the BIG-AFF project (the continuation of the MAMIPEC project), either by other members of the aDeNu research group or by the partners on University of Valencia. These future works should take into account the issues discussed in Section 8.

11.1. Intra-subject approach

One of the most priority steps to take to go further in this research is the evaluation of the approach here presented following an intra-subject approach. This step requires the design of a long term experiment with participants, designing both long continuous experimental sessions as well as experimental sessions in different days for the same participants. By mean of doing this, some open issues from this work should be studied, such as the time validity of the model and the baseline (as well as new baseline possible approaches, to be described in section 11.2). Following this intra-subject approach would also allow us to compare the results from the inter-subject approach presented in this work and that intra-subject approach. From both approaches, the design of a combined model could also help to deal with some well-known modeling problems such as the cold start problem [217].

11.2. Interaction baseline model

One of the main issues researched in stage 2 was the use of a baseline in order to normalize the data collected from interaction devices. Although the proposed approach threw promising results in the use of the dynamic baseline in predicting valence values, further evaluations can be done. The evaluation of the proposed approach in this work within an inter-subject approach could help to strength the results obtained in this work. In addition, new baseline approaches and variables could be evaluated: i) the repetition of the fixed baseline with a given frequency (which would be another variable to evaluate); ii) the evaluation of the dynamic baseline in long-term tasks (which may require the definition of a time window shorter than the task to generate the dynamic baseline); iii) possible modifications to the baseline task in order to require more interaction from all the different interaction data sources.

11.3. Experimental environment-related limitations

Further experiments are needed in order to take the proposed approach to a real e-Learning scenario. Although in this work we took the approach here introduced to a real classroom, another goal scenario would be the use of the approach here presented in a system to be used at the participants' houses with no need of intervention by an observer. This issue involves changes in many aspects, from the data sources (as it would be interesting the use of currently publicly available wearable devices) to the data privacy and the setup needed to be done by the participants.

11.4. Data preparation

Regarding the data preparation, there is still a lot of work to be done in order to carry out a precise evaluation of the impact the different data preparation approaches possible. When analyzing the different possibilities of the model generation from the data here proposed, the possible approaches to be carried out are countless. As we saw in section 2.2, there are a huge number of model creation techniques as well as data preprocessing steps (so many different approaches can be generated form combining them).

11.5. Affective state representation

Other open issue identified in this field is how to model or represent the affective states of the user. In both stages of this work a dimensional representation has been used, while in the ITS experiment, a categorical approach has been used. Nevertheless, no comparison between both approaches has been done. It has been seen in section 2.3.1 how there are many works following both approaches but no comparison between results using both approaches has been done.

Additionally, in both stages, the emotional labels used were discretized, which raises the issue of the discretization approach. Although that issue has been partially addressed in stage 2, many different approaches are possible regarding that methodological variable. In that sense, those many approaches could be evaluated, but in this sense, the further evaluation of the discretization approach should be linked to the future work described in section 11.11. The design of the implementation of the AMO-ML methodology here described in a recommender system, would force the emotional representation required by the recommendations to be triggered.

11.6. Data sources

Although in a closed set of data sources has been used in this work, reviewing the related works makes us realize that there are a wide range of data sources to use in addition to the ones proposed here (see section 2.1). The addition of new data sources could enrich the model of the user, providing some pros and cons to the new approach generated through that. In the pros side, we can find new ways to detect affective reactions of the participants. The proposed AMO-ML methodology should be ready to handle the inclusion of new data sources as the use of feature selection techniques will select only those features providing more information (even following an intra-subject approach could include the feature of suggesting the users get rid of those data sources that are less informative to the system, aiming to an even less intrusive approach). The inclusion of new data sources could also help to model those participants that cannot use some of the data sources proposed (as explained in section 11.10).

When implementing the AMO-ML methodology here proposed in a system, another issue to take into account in this sense is not the inclusion of data sources that measure other signals not already recorded by the current data sources, but the abstraction from the device used to measure each signal. In this work, the devices used (especially the physiological signal recording ones) have had a strong impact on some methodological issues presented. The idea in future works is the development of an interface that makes the system sonly aware of signals but not of devices. In the future, this layer would help to make the system proposed compatible with many devices, requiring only the implementation of the connection between each device and that interface connected to the system.

11.7. Time window

In the experiments celebrated during the development of this research work, the time windows used in both stages have been delimited by the task performed by the participant. In case a long-term intra-subject approach is evaluated in future works (as described in section 11.1), other temporal approaches seen in related works (such as fixed time window) could be evaluated and compared. Nevertheless, the decision taken in the experiments was driven by the nature of the task proposed, as they were short tasks that provided a short time window to be modeled. As seen in section 2.3.3.a, in open world experiments, it is common to use fixed time windows. That approach has been used in the ITS experiment, evaluating the best fixed time window to be used, but no comparison between both approaches (fixed time window and task delimited labeling) has been carried out. In further experiments that would be an interesting point of evaluation. Other interesting point that has not been analyzed is the possibility of carry out an experiment with a task-delimited labeling with multiple duration tasks, as all the tasks proposed In the experiments here described had the same duration (within the same experiment).

11.8. Live data processing and data synchronization

Another direction this research could be driven to is the optimization of the technical processes of the model generation. Although the goal of this work is the proposal of the AMO-ML methodology to follow in order to provide affective state detection from a combination of data sources using machine learning techniques, the approach followed in this goal did not provide that functionality in real time. This limitation has been present due to the limitations of the hardware or the software used, but what here has been introduced, could be implemented in order to provide a real-time modeling of the participant. The first step to take in that direction would be the implementation of the model generation system to be compatible with different data gathering APIs. Once that has been done, different issues should be resolved, from the time window to be used (as the use of some data preprocessing or model generation techniques might induce a big delay in the model generation) to the computational requirements to provide live modeling capabilities to a system (with a given number of users).

Obviously, that kind of capabilities, require that some other issues already faced in the work here presented to be resolved. The synchronization of the data has been an open issue faced in stage 1 and solved in stage 2. In case the approach of the previous paragraph is to be carried out, the inclusion of a server that receives the data live could help to mitigate the synchronization problem (supposing all the signals are sent live to the server), although the transmission of the data might also induce data loss and some data privacy problems.

11.9. Elicitation methods

Regarding the elicitation methods used in this work, the same strategies have been used in the three experiments carried out. This similarity is due to the constraints the educational field imposes in the possible elicitation methods identified. In section 2.4, many different elicitation methods were found in related works, but in order not to make the participants to leave the educational context, some of them were discarded, setting our focus on those elicitators that could take part in a normal educational scenario (e.g. different difficulty levels, time limits, etc.). Due to the nature of the tasks proposed in the experiments, no other elicitators were used, but different educational tasks could be proposed. For example, the inclusion of videos could be in order to explain some concepts or songs in Music learning. In art related classes, the choice of the artworks to show could also induce many emotions (and that could be used to choose the materials that could lead the participants to some desired affective states).

11.10. Accessibility

The results here presented belong to experiments where there were no participants with special needs. Nevertheless, we cannot assume all the potential users of a system based in the research here presented will not have any special need. In this sense, although some experiments were carried out with people with visual impairments [211], it is required new experiments in order to further evaluate the application (and possible adaptations) of the proposed approach. More experiments in detecting affective states are also required with people with other kind of disabilities, going from motor skills problems (which might impact on the data recorded from some of the proposed data sources in this works, such as keyboard or mouse) to mental illnesses. Is in the mental illnesses field where we could find a wide variety of different profiles to take into account in a very special and detailed way. We could find people with mood disorders, with very complex behaviors to people with cognitive disorders. All these issues require a huge work behind, first, in order to try to make the system capable to identify some of these profiles, and second, in order to model that in a special way. The modeling of these kind of profiles, is an issue that conveys a lot of work, putting especial effort on things such as the model privacy or making sure that the system where the emotional model is integrated, has the proper strategies to deal with these kind of profiles implemented.

Another issue to take into account related to the accessibility is the requirement to make accessible the interfaces of all the developed tools related to this work. This requirement should not only be carried out in the tools to be used by participants (such as MOKEETO), but also in the tools to be used by labelers.

11.11. **Affective model use**

One of the most sensible next steps to take is the integration of the proposed model generation approach into a recommender system. This future work is one of the key things to do in order to evaluate its applicability in real contexts. By means of performing this integration, it is predictable that a big number of new issues will appear from this integration. From the most suitable way to model the affective state (categorical or dimensional approach, and, in case of the dimensional approach, numerical or discretized approach), to more technical issues related to the frequency of the model generation. This last issue is also quite related to the development of intra-subject experiments (as mentioned in section 11.1).

11.12. **Dissemination**

As it can be seen in Appendix I (section 13.1), there are many publications (in conferences, journals, etc.) that have been generated from the work here presented. Nevertheless, some other papers are to be presented in order to reflect the outcomes here presented. The last published paper prior to the defense of this thesis has been [193] which reflects partially the work carried out in stage 2, and other work has been reviewed and is being polished in order to get published in the Transactions on Intelligent Systems and Technology journal. Other materials from this work are to be published.

12. References

1. A. Ahmed Ahmed and Issa Traore. 2014. Biometric Recognition Based on Free-Text Keystroke Dynamics. *IEEE Transactions on Cybernetics* 44, 4: 458–472. https://doi.org/10.1109/TCYB.2013.2257745
2. Mary Ainley, Matthew Corrigan, and Nicholas Richardson. 2005. Students, tasks and emotions: Identifying the contribution of emotions to students' reading of popular culture and popular science texts. *Learning and Instruction* 15, 5: 433–447. https://doi.org/10.1016/j.learninstruc.2005.07.011
3. Soraia M. Alarcao and Manuel J. Fonseca. 2017. Emotions Recognition Using EEG Signals: A Survey. *IEEE Transactions on Affective Computing* PP, 99: 1–1. https://doi.org/10.1109/TAFFC.2017.2714671
4. Aseel AL-Ayash, Robert T. Kane, Dianne Smith, and Paul Green-Armytage. 2016. The influence of color on student emotion, heart rate, and performance in learning environments. *Color Research & Application* 41, 2: 196–205. https://doi.org/10.1002/col.21949
5. Md Liakat Ali, John V. Monaco, Charles C. Tappert, and Meikang Qiu. 2017. Keystroke Biometric Systems for User Authentication. *Journal of Signal Processing Systems* 86, 2–3: 175–190. https://doi.org/10.1007/s11265-016-1114-9
6. Juan Miguel L. Andres and Ma Mercedes T. Rodrigo. 2015. Analyzing Student Action Sequences and Affect While Playing Physics Playground. In *Proceedings of the Workshops at the 17th International Conference on Artificial Intelligence in Education, AIED 2015, Madrid, Spain, June 22 + 26, 2015* (CEUR Workshop Proceedings), 24–33. Retrieved March 13, 2017 from http://ceur-ws.org/Vol-1432/amadl_pap4.pdf
7. Miguel Arevalillo-Herráez and David Arnau. 2013. A Hypergraph Based Framework for Intelligent Tutoring of Algebraic Reasoning. In *Artificial Intelligence in Education*, H. Chad Lane, Kalina Yacef, Jack Mostow and Philip Pavlik (eds.). Springer Berlin Heidelberg, 512–521. Retrieved October 4, 2013 from http://link.springer.com/chapter/10.1007/978-3-642-39112-5_52
8. Miguel Arevalillo-Herráez, David Arnau, Jesus G. Boticario, José Antonio González-Calero, Paloma Moreno-Clari, Salvador Moreno-Picot, Sergio Salmeron-Majadas, and Olga C. Santos. 2014. Computación afectiva: desarrollos y propuestas de uso en la construcción de sistemas de tutorización inteligente.
9. Miguel Arevalillo-Herráez, David Arnau, and Luis Marco-Giménez. 2013. Domain-specific knowledge representation and inference engine for an intelligent tutoring system. *Knowledge-Based Systems* 49: 97–105. https://doi.org/10.1016/j.knosys.2013.04.017
10. Miguel Arevalillo-Herráez, David Arnau, Luis Marco-Giménez, José A. González-Calero, Salvador Moreno-Picot, Paloma Moreno-Clari, Aladdin Ayesh, Olga C. Santos, Jesús G. Boticario, Mar Saneiro, Sergio Salmeron-Majadas, Pilar Quirós, and Raúl Cabestrero. 2014. Providing Personalized Guidance in Arithmetic Problem Solving. In *Personalization Approaches in Learning Environments*, 42–

48. Retrieved September 5, 2014 from http://ceur-ws.org/Vol-1181/pale2014_paper_05.pdf

11. Miguel Arevalillo-herráez, Salvador Moreno-picot, David Arnau, Paloma Moreno, Jesus G. Boticario, Olga C. Santos, Raúl Cabestrero, Pilar Quirós, Sergio Salmeron-Majadas, Ángeles Manjarrés-riesco, and Mar Saneiro. 2013. Towards Enriching an ITS with Affective Support. In *Personalization Approaches in Learning Environments*, 5–12. Retrieved from http://ceur-ws.org/Vol-997/pale2013_paper_1.pdf
12. David Arnau, Miguel Arevalillo-Herráez, Luis Puig, and José Antonio González-Calero. 2013. Fundamentals of the design and the operation of an intelligent tutoring system for the learning of the arithmetical and algebraic way of solving word problems. *Computers & Education* 63: 119–130. https://doi.org/10.1016/j.compedu.2012.11.020
13. Ivon Arroyo, Beverly Park Woolf, Winslow Burelson, Kasia Muldner, Dovan Rai, and Minghui Tai. 2014. A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect. *International Journal of Artificial Intelligence in Education* 24, 4: 387–426. https://doi.org/10.1007/s40593-014-0023-y
14. Andreia Artífice, Fernando Ferreira, Elsa Marcelino-Jesus, João Sarraipa, and Ricardo Jardim-Gonçalves. 2017. Student’s Attention Improvement Supported by Physiological Measurements Analysis. In *Technological Innovation for Smart Systems* (IFIP Advances in Information and Communication Technology), 93–102. https://doi.org/10.1007/978-3-319-56077-9_8
15. Mark H. Ashcraft and Elizabeth P. Kirk. 2001. The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General* 130, 2: 224–237. https://doi.org/10.1037/0096-3445.130.2.224
16. Mark H. Ashcraft and Jeremy A. Krause. 2007. Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review* 14, 2: 243–248. https://doi.org/10.3758/BF03194059
17. Sinem Aslan, Sinem Emine Mete, Eda Okur, Ece Oktay, Nese Alyuz, Utku Ergin Genc, David Stanhill, and Asli Arslan Esme. 2017. Human Expert Labeling Process (HELP): Towards a Reliable Higher-order User State Labeling Process and Tool to Assess Student Engagement. *Educational technology: The magazine for managers of change in education* 57, 1: 53–59.
18. Aladdin Ayesh, Miguel Arevalillo-Herraez, and Francesc .J. Ferri. 2014. Cognitive reasoning and inferences through psychologically based personalised modelling of emotions using associative classifiers. In *2014 IEEE 13th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, 67–72. https://doi.org/10.1109/ICCI-CC.2014.6921443
19. Judith Azcarraga and Merlin Teodosia Suarez. 2013. Recognizing Student Emotions using Brainwaves and Mouse Behavior Data: *International Journal of Distance Education Technologies* 11, 2: 1–15. https://doi.org/10.4018/jdet.2013040101
20. Ryan S. J. d. Baker, Sidney K. D’Mello, Ma.Mercedes T. Rodrigo, and Arthur C. Graesser. 2010. Better to Be Frustrated Than Bored: The Incidence, Persistence, and Impact of Learners’ Cognitive-affective States During Interactions with Three Different Computer-based Learning Environments. *Int. J. Hum.-Comput. Stud.* 68, 4: 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003
21. Kaveh Bakhtiyari, Mona Taghavi, and Hafizah Husain. 2015. Hybrid affective computing—keyboard, mouse and touch screen: from review to experiment.

Neural Computing and Applications 26, 6: 1277–1296. https://doi.org/10.1007/s00521-014-1790-y

22. Bengi Baran, Edward F. Pace-Schott, Callie Ericson, and Rebecca M. C. Spencer. 2012. Processing of Emotional Reactivity and Emotional Memory over Sleep. *The Journal of Neuroscience* 32, 3: 1035–1042. https://doi.org/10.1523/JNEUROSCI.2532-11.2012
23. Lisa Feldman Barrett. 1998. Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus. *Cognition & Emotion* 12, 4: 579–599. https://doi.org/10.1080/026999398379574
24. Veronica Benet-Martinez and Oliver P. John. 1998. Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *Journal of personality and social psychology* 75, 3: 729.
25. Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. 2008. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*, Christine Preisach, Professor Dr Hans Burkhardt, Professor Dr Lars Schmidt-Thieme and Professor Dr Reinhold Decker (eds.). Springer Berlin Heidelberg, 319–326. Retrieved February 20, 2014 from http://link.springer.com/chapter/10.1007/978-3-540-78246-9_38
26. Samit Bhattacharya. 2017. A Predictive Linear Regression Model for Affective State Detection of Mobile Touch Screen Users. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 9, 1: 30–44. https://doi.org/10.4018/IJMHCI.2017010103
27. Robert Bixler and Sidney D'Mello. 2013. Detecting Boredom and Engagement During Writing with Keystroke Analysis, Task Appraisals, and Stable Traits. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (IUI '13), 225–234. https://doi.org/10.1145/2449396.2449426
28. Robert Bixler and Sidney D'Mello. 2013. Towards Automated Detection and Regulation of Affective States During Academic Writing. In *Artificial Intelligence in Education*, H. Chad Lane, Kalina Yacef, Jack Mostow and Philip Pavlik (eds.). Springer Berlin Heidelberg, 904–907. Retrieved October 4, 2013 from http://link.springer.com/chapter/10.1007/978-3-642-39112-5_142
29. Susana Bloch, Madeleine Lemeignan, and Nancy Aguilera-T. 1991. Specific respiratory patterns distinguish among human basic emotions. *International Journal of Psychophysiology* 11, 2: 141–154. https://doi.org/10.1016/0167-8760(91)90006-J
30. Erik Boiy, Pieter Hens, Koen Deschacht, and Marie-francine Moens. 2007. Automatic sentiment analysis in on-line text. In *In Proceedings of the 11th International Conference on Electronic Publishing*, 349–360.
31. Nigel Bosch, Huili Chen, Sidney D'Mello, Ryan Baker, and Valerie Shute. 2015. Accuracy vs. Availability Heuristic in Multimodal Affect Detection in the Wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (ICMI '15), 267–274. https://doi.org/10.1145/2818346.2820739
32. Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2015. Automatic Detection of Learning-Centered Affective States in the Wild. In *Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015). ACM, New York, NY, USA*. Retrieved February 23, 2015 from http://myweb.fsu.edu/vshute/pdf/bosch.pdf

33. Anthony F. Botelho, Ryan S. Baker, and Neil T. Heffernan. 2017. Improving Sensor-Free Affect Detection Using Deep Learning. In *Artificial Intelligence in Education* (Lecture Notes in Computer Science), 40–51. https://doi.org/10.1007/978-3-319-61425-0_4

34. Jesus G. Boticario, Olga C. Santos, Raúl Cabestrero, Pilar Quirós, Sergio Salmerón-Majadas, Raúl Uria-Rivas, Mar Saneiro, Miguel Arevalillo-Herráez, and Francesc J. Ferri. 2017. BIG-AFF: Exploring Low Cost and Low Intrusive Infrastructures for Affective Computing in Secondary Schools. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (UMAP '17), 287–292. https://doi.org/10.1145/3099023.3099084

35. Jesus G. Boticario, Olga C. Santos, Raúl Cabestrero, Pilar Quirós, Mar Saneiro, Sergio Salmeron-Majadas, Ángeles Manjarrés, Alejandro Rodriguez-Ascaso, Elena del Campo, and Emmanuelle Raffene. 2014. Avances en el modelado de aspectos afectivos en escenarios educativos inclusivos y personalizados. In *VII Jornadas de Redes de Investigación en Innovación Docente de la UNED*.

36. Beverly L. Bower and Kimberly P. Hardy. 2004. From correspondence to cyberspace: Changes and challenges in distance education. *New Directions for Community Colleges* 2004, 128: 5–12. https://doi.org/10.1002/cc.169

37. Elaine Boyd and Norman Whitby. 2009. *Activate! B2 Use of English*. Pearson Longman.

38. Marc A. Brackett, Raquel Palomera, Justyna Mojsa-Kaja, Maria Regina Reyes, and Peter Salovey. 2010. Emotion-regulation ability, burnout, and job satisfaction among British secondary-school teachers. *Psychology in the Schools* 47, 4: 406–417. https://doi.org/10.1002/pits.20478

39. Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1: 49–59.

40. Margaret M. Bradley and Peter J. Lang. 2007. The International Affective Digitized Sounds (IADS-2): Affective ratings of sounds and instruction manual. *Center for Research in Psychophysiology, Gainesville, FL, Tech. Rep. B-3*. Retrieved September 3, 2014 from ftp://flash.ict.usc.edu/arizzo/IADs%20Audio/IADS2%20Tech%20Report/IADS2.pdf

41. Leo Breiman. 1996. Bagging Predictors. *Machine Learning* 24, 2: 123–140. https://doi.org/10.1023/A:1018054314350

42. David Guy Brizan, Adam Goodkind, Patrick Koch, Kiran Balagani, Vir V. Phoha, and Andrew Rosenberg. 2015. Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies* 82: 57–68. https://doi.org/10.1016/j.ijhcs.2015.04.005

43. Egon L. van den Broek. 2013. Ubiquitous emotion-aware computing. *Personal and Ubiquitous Computing* 17, 1: 53–67.

44. Egon L. van den Broek, Viliam Lisỳ, Joris H. Janssen, Joyce HDM Westerink, Marleen H. Schut, and Kees Tuinenbreijer. 2010. Affective man-machine interface: unveiling human emotions through biosignals. In *Biomedical Engineering Systems and Technologies*. Springer, 21–47. Retrieved November 20, 2013 from http://link.springer.com/chapter/10.1007/978-3-642-11721-3_2

45. Richard P. Brown, Patricia L. Gerbarg, and Fred Muench. 2013. Breathing Practices for Treatment of Psychiatric and Stress-Related Medical Conditions. *Psychiatric Clinics of North America* 36, 1: 121–140. https://doi.org/10.1016/j.psc.2013.01.001

46. Tadeus Caliński and Joachim Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics* 3, 1: 1–27. https://doi.org/10.1080/03610927408827101
47. Erik Cambria, Paolo Gastaldo, Federica Bisio, and Rodolfo Zunino. 2014. An ELM-based model for affective analogical reasoning. *Neurocomputing* 1: 18.
48. S. Campanella, G. Dimauro, A. Ferrante, D. Impedovo, S. Impedovo, M. G. Lucchese, R. Modugno, G. Pirlo, L. Sarcinella, E. Stasolla, and others. 2008. E-learning platforms in the Italian Universities: the technological solutions at the University of Bari. *WSEAS Transactions on Advances in Engineering Education* 5, 1: 12–19.
49. Alexandra Cernian, Adriana Olteanu, Dorin Carstoiu, and Cristina Mares. 2017. Mood Detector - On Using Machine Learning to Identify Moods and Emotions. In *2017 21st International Conference on Control Systems and Computer Science (CSCS)*, 213–216. https://doi.org/10.1109/CSCS.2017.36
50. Nitesh V. Chawla. 2009. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*. Springer, Boston, MA, 875–886. https://doi.org/10.1007/978-0-387-09823-4_45
51. Jingjing Chen, Bin Zhu, Olle Balter, Jianliang Xu, Weiwen Zou, Anders Hedman, Rongchao Chen, and Mengdie Sang. 2017. FishBuddy: Promoting Student Engagement in Self-Paced Learning through Wearable Sensing. In *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*, 1–9. https://doi.org/10.1109/SMARTCOMP.2017.7947008
52. Ming-Syan Chen, Jiawei Han, and P.S. Yu. 1996. Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering* 8, 6: 866–883. https://doi.org/10.1109/69.553155
53. Kwang-Ho Choi, Junbeom Kim, O. Sang Kwon, Min Ji Kim, Yeon Hee Ryu, and Ji-Eun Park. 2017. Is heart rate variability (HRV) an adequate tool for evaluating human emotions? – A focus on the use of the International Affective Picture System (IAPS). *Psychiatry Research* 251: 192–196. https://doi.org/10.1016/j.psychres.2017.02.025
54. Christian Collet, Evelyne Vernet-Maury, Georges Delhomme, and André Dittmar. 1997. Autonomic nervous system response patterns specificity to basic emotions. *Journal of the Autonomic Nervous System* 62, 1: 45–57. https://doi.org/10.1016/S0165-1838(96)00108-7
55. Daniel T. Cordaro, Rui Sun, Dacher Keltner, Shanmukh Kamble, Niranjan Huddar, and Galen McNeil. 2018. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion (Washington, D.C.)* 18, 1: 75–93. https://doi.org/10.1037/emo0000302
56. Shaundra B. Daily, Melva T. James, David Cherry, John J. Porter III, Shelby S. Darnell, Joseph Isaac, and Tania Roy. 2017. Affective Computing: Historical Foundations, Current Applications, and Future Trends. In *Emotions and Affect in Human Factors and Human-Computer Interaction*, Myounghoon Jeon (ed.). Academic Press, San Diego, 213–231. https://doi.org/10.1016/B978-0-12-801851-4.00009-4
57. Keya Das, Sarwar Ali, Koyo Otsu, Hisato Fukuda, Antony Lam, Yoshinori Kobayashi, and Yoshinori Kuno. 2017. Detecting Inner Emotions from Video Based Heart Rate Sensing. In *Intelligent Computing Methodologies* (Lecture Notes in Computer Science), 48–57. https://doi.org/10.1007/978-3-319-63315-2_5
58. Valerie A. DeBellis and Gerald A. Goldin. 1997. The affective domain in mathematical problem solving. In *Proceedings of the 21st Conference of the*

International Group for the Psychology of Mathematics Education, Vol. 2, 209–216.

59. Matt Dennis, Judith Masthoff, and Chris Mellish. 2012. Towards a model of Personality, Affective State, Feedback and Learner Motivation. Retrieved April 22, 2013 from http://ceur-ws.org/Vol-872/pale2012_paper_3.pdf
60. Francesca D'Errico, Marinella Paciello, and Luca Cerniglia. 2016. When emotions enhance students' engagement in e-learning processes. *Journal of e-Learning and Knowledge Society* 12, 4. Retrieved November 28, 2017 from https://www.learntechlib.org/p/173676/
61. Ilse van Diest, Winnie Winters, Stephan Devriese, Elke Vercamst, Jiang N. Han, Karel P. van de Woestijne, and Omer van den Bergh. 2001. Hyperventilation beyond fight/flight: Respiratory responses during emotional imagery. *Psychophysiology* 38, 6: 961–968. https://doi.org/10.1111/1469-8986.3860961
62. Sidney D'Mello and Art Graesser. 2010. Mining Bodily Patterns of affective experience during learning. In *Proceedings of the 3rd International Conference on Educational Data Mining*, 31–40. Retrieved February 25, 2013 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.174.6434&rep=rep1&type=pdf
63. Sidney K. D'Mello. 2014. Emotional Rollercoasters: Day Differences in Affect Incidence during Learning. In *The Twenty-Seventh International Flairs Conference*. Retrieved July 30, 2014 from http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS14/paper/download/7777/7875
64. Sidney K. D'Mello. 2016. On the Influence of an Iterative Affect Annotation Approach on Inter-Observer and Self-Observer Reliability. *IEEE Transactions on Affective Computing* 7, 2: 136–149. https://doi.org/10.1109/TAFFC.2015.2457413
65. Sidney K. D'Mello and Jacqueline Kory. 2015. A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Comput. Surv.* 47, 3: 43:1–43:36. https://doi.org/10.1145/2682899
66. Peter Dolog, Nicola Henze, Wolfgang Nejdl, and Michael Sintek. 2004. Personalization in Distributed e-Learning Environments. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers &Amp; Posters* (WWW Alt. '04), 170–179. https://doi.org/10.1145/1013367.1013395
67. Marieke van Dooren, J. J. G. de Vries, and Joris H. Janssen. 2012. Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & behavior* 106, 2: 298–304.
68. Joseph A. Durlak, Roger P. Weissberg, Allison B. Dymnicki, Rebecca D. Taylor, and Kriston B. Schellinger. 2011. The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development* 82, 1: 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x
69. Matthieu Duvinage, Thierry Castermans, Thierry Dutoit, Mathieu Petieau, Thomas Hoellinger, Caty De Saedeleer, Karthik Seetharaman, and Guy Cheron. 2012. A P300-based Quantitative Comparison between the Emotiv Epoc Headset and a Medical EEG Device. https://doi.org/10.2316/P.2012.764-071
70. Tuomas Eerola and Jonna K. Vuoskoski. 2010. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*. https://doi.org/10.1177/0305735610362821

71. Mohamad A. Eid and Hussein Al Osman. 2016. Affective Haptics: Current Research and Future Directions. *IEEE Access* 4: 26–40. https://doi.org/10.1109/ACCESS.2015.2497316
72. Paul Ekman. 2016. What Scientists Who Study Emotion Agree About. *Perspectives on Psychological Science* 11, 1: 31–34. https://doi.org/10.1177/1745691615596992
73. Paul Ekman and Dacher Keltner. 1970. Universal facial expressions of emotion. *California Mental Health Research Digest* 8, 4: 151–158.
74. Paul Ekman, Robert W. Levenson, and Wallace V. Friesen. 1983. Autonomic nervous system activity distinguishes among emotions. *Science* 221, 4616: 1208–1210. https://doi.org/10.1126/science.6612338
75. Paul Ekman and Harriet Oster. 1979. Facial Expressions of Emotion. *Annual Review of Psychology* 30, 1: 527–554. https://doi.org/10.1146/annurev.ps.30.020179.002523
76. C. Epp, M. Lippold, and R. L. Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, 715–724.
77. Clayton Epp. 2010. Identifying emotional states through keystroke dynamics. University of Saskatchewan.
78. Clayton Epp, Michael Lippold, and Regan L. Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, 715–724. Retrieved February 25, 2013 from http://dl.acm.org/citation.cfm?id=1979046
79. Daniel A. M. Felipe, Kim I. N. Gutierrez, Eilynn C. M. Quiros, and Larry A. Vea. 2012. Towards the Development of Intelligent Agent for Novice C/C++ Programmers through Affective Analysis of Event Logs. *Proceedings of the International MultiConference of Engineers and Computer Scientists* 1: 511–518.
80. Lauren Fratamico, Cristina Conati, Samad Kardan, and Ido Roll. 2017. Applying a Framework for Student Modeling in Exploratory Learning Environments: Comparing Data Representation Granularity to Handle Environment Complexity. *International Journal of Artificial Intelligence in Education* 27, 2: 320–352. https://doi.org/10.1007/s40593-016-0131-y
81. Eugene Y. Fu, Tiffany C. K. Kwok, Erin Y. Wu, Hong V. Leong, Grace Ngai, and Stephen C. F. Chan. 2017. Your Mouse Reveals Your Next Activity: Towards Predicting User Intention from Mouse Interaction. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, 869–874. https://doi.org/10.1109/COMPSAC.2017.270
82. Yujun Fu, Hong Va Leong, Grace Ngai, Michael Xuelin Huang, and Stephen C. F. Chan. 2017. Physiological mouse: toward an emotion-aware mouse. *Universal Access in the Information Society* 16, 2: 365–379. https://doi.org/10.1007/s10209-016-0469-9
83. Ennio Gambi, Angela Agostinelli, Alberto Belli, Laura Burattini, Enea Cippitelli, Sandro Fioretti, Paola Pierleoni, Manola Ricciuti, Agnese Sbrollini, and Susanna Spinsante. 2017. Heart Rate Detection Using Microsoft Kinect: Validation and Comparison to Wearable Devices. *Sensors* 17, 8: 1776. https://doi.org/10.3390/s17081776
84. Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. 2012. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 4: 31.

85. Paul van Gent. Analyzing a Discrete Heart Rate Signal Using Python – 2nd Part. Retrieved May 10, 2018 from http://www.paulvangent.com/2016/03/21/analyzing-a-discrete-heart-rate-signal-using-python-part-2/
86. Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017. Evaluating Effectiveness of Smartphone Typing as an Indicator of User Emotion. In *Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 146–151. https://doi.org/10.1109/ACII.2017.8273592
87. Kiel Gilleade, Alan Dix, and Jen Allanson. 2005. Affective Videogames and Modes of Affective Gaming: Assist Me, Challenge Me, Emote Me. Retrieved February 21, 2014 from http://www.gamesconference.org/digra2005/viewabstract.php?id=256
88. Atefeh Goshvarpour, Ataollah Abbasi, Ateke Goshvarpour, and Sabalan Daneshvar. 2016. Fusion Framework for Emotional Electrocardiogram and Galvanic Skin Response Recognition: Applying Wavelet Transform. *Iranian Journal of Medical Physics* 13, 3: 163–173. https://doi.org/10.22038/ijmp.2016.7960
89. Marco Granato, Davide Gadia, Dario Maggiorini, and Laura Anna Ripamonti. 2017. Emotions Detection Through the Analysis of Physiological Information During Video Games Fruition. In *Games and Learning Alliance* (Lecture Notes in Computer Science), 197–207. https://doi.org/10.1007/978-3-319-71940-5_18
90. Shalom Greene, Himanshu Thapliyal, and Allison Caban-Holt. 2016. A Survey of Affective Computing for Stress Detection: Evaluating technologies in stress detection for better health. *IEEE Consumer Electronics Magazine* 5, 4: 44–56. https://doi.org/10.1109/MCE.2016.2590178
91. James J. Gross and Ross A. Thompson. 2007. Emotion regulation: Conceptual foundations. *Handbook of emotion regulation* 3: 24.
92. Daniele Gunetti and Claudia Picardi. 2005. Keystroke Analysis of Free Text. *ACM Trans. Inf. Syst. Secur.* 8, 3: 312–347. https://doi.org/10.1145/1085126.1085129
93. Isabelle Guyon and André Elisseeff. 2003. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* 3: 1157–1182.
94. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1: 10–18. https://doi.org/10.1145/1656274.1656278
95. Hamza Hamdi, Paul Richard, Aymeric Suteau, and Philippe Allain. 2012. Emotion assessment for affective computing based on physiological responses. In *2012 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8. https://doi.org/10.1109/FUZZ-IEEE.2012.6250778
96. Santoso Handri, Kuniaki Yajima, Shusaku Nomura, Nobuyuki Ogawa, Yoshimasa Kurosawa, and Yoshimi Fukumura. 2010. Evaluation of Student’s Physiological Response Towards E-Learning Courses Material by Using GSR Sensor. In *Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on*, 805–810. Retrieved November 20, 2013 from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5591037
97. Katrin Hänsel, Akram Alomainy, and Hamed Haddadi. 2016. Large Scale Mood and Stress Self-assessments on a Smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (UbiComp ’16), 1180–1184. https://doi.org/10.1145/2968219.2968305

98. Linda Harasim. 2000. Shift happens: online education as a new paradigm in learning. *The Internet and Higher Education* 3, 1–2: 41–61. https://doi.org/10.1016/S1096-7516(00)00032-4
99. Jason M. Harley, François Bouchet, M. Sazzad Hussain, Roger Azevedo, and Rafael Calvo. 2015. A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior* 48: 615–625. https://doi.org/10.1016/j.chb.2015.02.013
100. Martin Hibbeln, Jeffrey L. Jenkins, Christoph Schneider, JosephS Valacich, and Markus Weinmann. 2017. How is Your User Feeling? Inferring Emotion Through Human-Computer Interaction Devices. *MIS Quarterly: Management Information Systems* 41, 1: 1–22.
101. Martin Thomas Hibbeln, Jeffrey L. Jenkins, Christoph Schneider, Joseph Valacich, and Markus Weinmann. 2016. *Inferring Negative Emotion from Mouse Cursor Movements*. Social Science Research Network, Rochester, NY. Retrieved November 30, 2016 from https://papers.ssrn.com/abstract=2708108
102. H. Hinrichs and Wielant Machleidt. 1992. Basic emotions reflected in EEG-coherences. *International Journal of Psychophysiology* 13, 3: 225–232. https://doi.org/10.1016/0167-8760(92)90072-J
103. Markus Hofmann and Ralf Klinkenberg (eds.). 2014. *RapidMiner: data mining use cases and business analytics applications*. CRC Press, Boca Raton.
104. Derek M. Isaacowitz, Heather A. Wadlinger, Deborah Goren, and Hugh R. Wilson. 2006. Selective preference in visual fixation away from negative images in old age? An eye-tracking study. *Psychology and Aging* 21, 1: 40–48. https://doi.org/10.1037/0882-7974.21.1.40
105. Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. 2000. Statistical Pattern Recognition: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 1: 4–37. https://doi.org/10.1109/34.824819
106. Lutz Jäncke. 1994. An EMG investigation of the coactivation of facial muscles during the presentation of affect-laden stimuli. *Journal of Psychophysiology* 8, 1: 1–10.
107. Natasha Jaques, Cristina Conati, Jason M. Harley, and Roger Azevedo. 2014. Predicting Affect from Gaze Data during Interaction with an Intelligent Tutoring System. In *Intelligent Tutoring Systems*, 29–38. https://doi.org/10.1007/978-3-319-07221-0_4
108. Frahang Jaryani, Shamsul Sahibudin, Mazdak Zamani, Maslin Masrom, Samaneh Salehy, and Jamshid Jamshidi. 2013. E-Learning Effects on Learning Quality: a Case Study of Iranian Students. *Journal of Administration and Development, Mahasarakham University วารสาร การ บริหาร และ พัฒนา* 1, 3: 16–27.
109. Jesus G. Boticario, Olga C. Santos, Raúl Cabestrero, Pilar QUirós, Mar Saneiro, Sergio Salmeron-Majadas, Ángeles Manjarrés, Alejandro Rodríguez Ascaso, Elena del Campo, and Emmanuelle Raffenne. 2013. Hacia el modelado de aspectos afectivos en escenarios educativos inclusivos y personalizados. In *VI Jornadas de Redes de Investigación en Innovación Docente de la UNED.*
110. George H. John and Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (UAI'95), 338–345. Retrieved February 20, 2014 from http://dl.acm.org/citation.cfm?id=2074158.2074196
111. John S. Noffsinger. 1926. *Correspondence Schools Lyceums Chautauquas*. The Macmillan Company. Retrieved September 1, 2014 from http://archive.org/details/correspondencesc028298mbp

112. William F. Johnson, Robert N. Emde, Klaus R. Scherer, and Mary D. Klinnert. 1986. Recognition of Emotion From Vocal Cues. *Archives of General Psychiatry* 43, 3: 280–283. https://doi.org/10.1001/archpsyc.1986.01800030098011
113. Imène Jraidi, Maher Chaouachi, and Claude Frasson. 2014. A Hierarchical Probabilistic Framework for Recognizing Learners' Interaction Experience Trends and Emotions. *Advances in Human-Computer Interaction* 2014. https://doi.org/10.1155/2014/632630
114. Arturas Kaklauskas, Mindaugas Krutinis, and Mark Seniut. 2009. Biometric mouse intelligent system for student's emotional and examination process analysis. In *Advanced Learning Technologies, 2009. ICALT 2009. Ninth IEEE International Conference on*, 189–193.
115. Iftikhar Ahmed Khan, Willem-Paul Brinkman, and Robert Hierons. 2013. DEPRECATED-Towards estimating computer users' mood from interaction behaviour with keyboard and mouse. *Frontiers of Computer Science*: 1–12. https://doi.org/10.1007/s11704-013-2331-z
116. Iftikhar Ahmed Khan, Willem-Paul Brinkman, and Robert Hierons. 2013. Towards estimating computer users' mood from interaction behaviour with keyboard and mouse. *Frontiers of Computer Science* 7, 6: 943–954. https://doi.org/10.1007/s11704-013-2331-z
117. Iftikhar Ahmed Khan, Willem-Paul Brinkman, and Robert M. Hierons. 2008. Towards a Computer Interaction-Based Mood Measurement Instrument. *Proc. PPIG2008, ISBN*: 971–978.
118. ChanMin Kim, Seung Won Park, and Joe Cozart. 2014. Affective and motivational factors of learning in online mathematics courses. *British Journal of Educational Technology* 45, 1: 171–185. https://doi.org/10.1111/j.1467-8535.2012.01382.x
119. ChanMin Kim and Reinhard Pekrun. 2014. Emotions and Motivation in Learning and Performance. In *Handbook of Research on Educational Communications and Technology*, J. Michael Spector, M. David Merrill, Jan Elen and M. J. Bishop (eds.). Springer New York, 65–75. Retrieved March 5, 2014 from http://link.springer.com/chapter/10.1007/978-1-4614-3185-5_6
120. A. Kolakowska. 2013. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 The 6th International Conference on Human System Interaction (HSI)*, 548–555. https://doi.org/10.1109/HSI.2013.6577879
121. Agata Kołakowska. 2013. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 6th International Conference on Human System Interactions (HSI)*, 548–555. https://doi.org/10.1109/HSI.2013.6577879
122. Agata Kołakowska. 2015. Recognizing emotions on the basis of keystroke dynamics. In *2015 8th International Conference on Human System Interaction (HSI)*, 291–297. https://doi.org/10.1109/HSI.2015.7170682
123. Agata Kołakowska. 2018. Usefulness of Keystroke Dynamics Features in User Authentication and Emotion Recognition. In *Human-Computer Systems Interaction*. Springer, Cham, 42–52. https://doi.org/10.1007/978-3-319-62120-3_4
124. Barry Kort, Rob Reilly, and Rosalind W. Picard. 2001. An Affective Model of Interplay between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion. In *Proceedings IEEE International Conference on Advanced Learning Technologies*, 43–46. https://doi.org/10.1109/ICALT.2001.943850

125. Davor Kukolja, Siniša Popović, Marko Horvat, Bernard Kovač, and Krešimir Ćosić. 2014. Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications. *International Journal of Human-Computer Studies* 72, 10: 717–727. https://doi.org/10.1016/j.ijhcs.2014.05.006
126. Pardis Lali, Maryam Naghizadeh, Hossein Nasrollahi, Hadi Moradi, and Maryam Mirian. 2014. Your mouse can tell about your emotions. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, 47–51. https://doi.org/10.1109/ICCKE.2014.6993360
127. Peter J. Lang, Margaret M. Bradley, and Bruce N. Cuthbert. 1999. International affective picture system (IAPS): Instruction manual and affective ratings. *The Center for Research in Psychophysiology, University of Florida*.
128. Raija M. T. Laukkanen and Paula K. Virtanen. 1998. Heart rate monitors: State of the art. *Journal of Sports Sciences* 16, sup1: 3–7. https://doi.org/10.1080/026404198366920
129. Hooseok Lee, Hoon Ko, Changwon Jeong, and Jinseok Lee. 2017. Wearable Photoplethysmographic Sensor Based on Different LED Light Intensities. *IEEE Sensors Journal* 17, 3: 587–588. https://doi.org/10.1109/JSEN.2016.2633575
130. Hosub Lee, Young Sang Choi, Sunjae Lee, and I. P. Park. 2012. Towards unobtrusive emotion recognition for affective social communication. In *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*, 260–264. Retrieved January 24, 2014 from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6181098
131. Po-Ming Lee, Wei-Hsuan Tsui, and Tzu-Chien Hsiao. 2014. The influence of emotion on keyboard typing: an experimental study using visual stimuli. *BioMedical Engineering OnLine* 13, 1: 81. https://doi.org/10.1186/1475-925X-13-81
132. Robert W. Levenson. 1992. Autonomic Nervous System Differences among Emotions. *Psychological Science* 3, 1: 23–27. https://doi.org/10.1111/j.1467-9280.1992.tb00251.x
133. Gavrielle Levine. 1996. Variability in Anxiety for Teaching Mathematics among Pre-service Elementary School Teachers Enrolled in a Mathematics Course. Retrieved September 2, 2014 from http://eric.ed.gov/?id=ED398067
134. Michael Lewis, Steven M. Alessandri, and Margaret W. Sullivan. 1992. Differences in Shame and Pride as a Function of Children’s Gender and Task Difficulty. *Child Development* 63, 3: 630–638. https://doi.org/10.1111/j.1467-8624.1992.tb01651.x
135. Xiao Li, Jessilyn Dunn, Denis Salins, Gao Zhou, Wenyu Zhou, Sophia Miryam Schüssler-Fiorenza Rose, Dalia Perelman, Elizabeth Colbert, Ryan Runge, Shannon Rego, Ria Sonecha, Somalee Datta, Tracey McLaughlin, and Michael P. Snyder. 2017. Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information. *PLOS Biology* 15, 1: e2001402. https://doi.org/10.1371/journal.pbio.2001402
136. Yee M. Lim, Aladdin Ayesh, and Martin Stacey. 2017. The effects of typing demand on learner’s Motivation/Attitude-driven Behaviour (MADB) model with mouse and keystroke behaviours. In *2017 IEEE 16th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, 175–181. https://doi.org/10.1109/ICCI-CC.2017.8109747
137. Yee Mei Lim, A. Ayesh, and M. Stacey. 2014. Detecting cognitive stress from keyboard and mouse dynamics during mental arithmetic. In *Science and*

Information Conference (SAI), 2014, 146–152. https://doi.org/10.1109/SAI.2014.6918183

138. Huan Liu and Hiroshi Motoda. 2012. *Feature Selection for Knowledge Discovery and Data Mining*. Springer Science & Business Media.
139. Jesús L. Lobo. 2014. La Reputación como indicador base para la agrupación de estudiantes en el Marco Lógico Colaborativo. Universidad Nacional de Educación a Distancia (UNED).
140. Hugo Lövheim. 2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses* 78, 2: 341–348. https://doi.org/10.1016/j.mehy.2011.11.016
141. Guohua Lu, F. Yang, J. A. Taylor, and John F. Stein. 2009. A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects. *Journal of Medical Engineering & Technology* 33, 8: 634–641. https://doi.org/10.3109/03091900903150998
142. Hai-Rong Lv, Zhong-Lin Lin, Wen-Jun Yin, and Jin Dong. 2008. Emotion recognition based on pressure sensor keyboards. In *2008 IEEE International Conference on Multimedia and Expo,* 1089–1092. https://doi.org/10.1109/ICME.2008.4607628
143. Ángeles Manjarrés-Riesco, Olga C. Santos, Jesus G. Boticario, and Mar Saneiro. 2013. Open Issues in Educational Affective Recommendations for Distance Learning Scenarios. In *CEUR Workshop Proceedings*, 26–33. Retrieved February 17, 2015 from http://ceur-ws.org/Vol-997/pale2013_paper_4.pdf
144. Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. 2017. Automatic Analysis of Facial Actions: A Survey. *IEEE Transactions on Affective Computing* PP, 99: 1–1. https://doi.org/10.1109/TAFFC.2017.2731763
145. M^{a} Antonia Manassero Más and ángel Vázquez Alonso. 1998. Validación de una Escala de Motivación de Logro. *Psicothema* 10, Número 2: 333–351.
146. Ebrahim Mazaheri, Marie-Odile Richard, and Michel Laroche. 2012. The role of emotions in online consumer behavior: a comparison of search, experience, and credence services. *Journal of Services Marketing* 26, 7: 535–550. https://doi.org/10.1108/08876041211266503
147. Rollin McCraty, Mike Atkinson, William A. Tiller, Glen Rein, and Alan D. Watkins. 1995. The effects of emotions on short-term power spectrum analysis of heart rate variability. *The American Journal of Cardiology* 76, 14: 1089–1093. https://doi.org/10.1016/S0002-9149(99)80309-9
148. Daniel McDuff, Sarah Gontarek, and Rosalind W. Picard. 2014. Remote measurement of cognitive stress via heart rate variability. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2957–2960. https://doi.org/10.1109/EMBC.2014.6944243
149. Richard A. McFarland. 1985. Relationship of skin temperature changes to the emotions accompanying music. *Biofeedback and Self-regulation* 10, 3: 255–267. https://doi.org/10.1007/BF00999346
150. Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4: 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011
151. Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology* 14, 4: 261–292. https://doi.org/10.1007/BF02686918

152. Helena M. Mentis and Geri Gay. 2002. Using TouchPad pressure to detect negative affect. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, 406–410. https://doi.org/10.1109/ICMI.2002.1167029
153. Batja Mesquita and Nico H. Frijda. 1992. Cultural variations in emotions: A review. *Psychological Bulletin* 112, 2: 179–204. https://doi.org/10.1037/0033-2909.112.2.179
154. Bishwas Mishra, Steven L. Fernandes, K. Abhishek, Aishwarya Alva, Chaithra Shetty, Chandan V. Ajila, Dhanush Shetty, Harshitha Rao, and Priyanka Shetty. 2015. Facial expression recognition using feature based techniques and model based techniques: A survey. In *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, 589–594. https://doi.org/10.1109/ECS.2015.7124976
155. Yehya Mohamad, Dirk T. Hettich, Elaina Bolinger, Niels Birbaumer, Wolfgang Rosenstiel, Martin Bogdan, and Tamara Matuz. 2014. Detection and Utilization of Emotional State for Disabled Users. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Deborah Fels, Dominique Archambault, Petr Peňáz and Wolfgang Zagler (eds.). Springer International Publishing, 248–255. Retrieved July 30, 2014 from http://link.springer.com/chapter/10.1007/978-3-319-08596-8_39
156. Michael G. Moore and Greg Kearsley. 2011. *Distance education: A systems view of online learning*. Cengage Learning. Retrieved September 1, 2014 from http://books.google.es/books?hl=es&lr=&id=dU8KAAAAQBAJ&oi=fnd&pg=PR4&dq=radio+tv+stations+educational+&ots=D1Xj_0Dsiw&sig=un4SJkNu8t6HQ51wFWHT5AkR_aY
157. Christian Mühl, Brendan Allison, Anton Nijholt, and Guillaume Chanel. 2014. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces* 1, 2: 66–84. https://doi.org/10.1080/2326263X.2014.912881
158. Peter K. Mungai and Runhe Huang. 2017. Using keystroke dynamics in a multi-level architecture to protect online examinations from impersonation. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*, 622–627. https://doi.org/10.1109/ICBDA.2017.8078710
159. A. F. M. Nazmul Haque Nahin, Jawad Mohammad Alam, Hasan Mahmud, and Kamrul Hasan. 2014. Identifying emotion by keystroke dynamics and text pattern analysis. *Behaviour & Information Technology* 33, 9: 987–996. https://doi.org/10.1080/0144929X.2014.907343
160. Eric C. Newburger. 1997. Computer use in the United States. *Population* 20: 522.
161. Domen Novak, Guillaume Chanel, Philippe Guillotel, and Alexander Koenig. 2017. Guest Editorial: Toward Commercial Applications of Affective Computing. *IEEE Transactions on Affective Computing* 8, 2: 145–147. https://doi.org/10.1109/TAFFC.2017.2676318
162. Jaclyn Ocumpaugh. 2015. Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences*.
163. Eileen O'Donnell, Séamus Lawless, Mary Sharp, and Vincent P. Wade. 2015. A Review of Personalised E-Learning: Towards Supporting Learner Diversity. *International Journal of Distance Education Technologies (IJDET)* 13, 1: 22–47. https://doi.org/10.4018/ijdet.2015010102

164. Arne Öhman, Francisco Esteves, Anders Flykt, and Joaquim J. F. Soares. 1993. Gateways to Consciousness: Emotion, Attention, and Electrodermal Activity. In *Progress in Electrodermal Research*. Springer, Boston, MA, 137–157. https://doi.org/10.1007/978-1-4615-2864-7_10
165. Kerry O'Regan. 2003. Emotion and e-learning. *Journal of Asynchronous learning networks* 7, 3: 78–92.
166. Charles E. Osgood. 1966. Dimensionality of the Semantic Space for Communication Via Facial Expressions. *Scandinavian Journal of Psychology* 7, 1: 1–30. https://doi.org/10.1111/j.1467-9450.1966.tb01334.x
167. Tom Page. 2015. Affective computing in the design of interactive systems. *i-manager's Journal on Mobile Applications and Technologies* 2, 2: 1–18.
168. Reinhard Pekrun. 1992. The Impact of Emotions on Learning and Achievement: Towards a Theory of Cognitive/Motivational Mediators. *Applied Psychology* 41, 4: 359–376. https://doi.org/10.1111/j.1464-0597.1992.tb00712.x
169. Reinhard Pekrun, Thomas Goetz, Anne C. Frenzel, Petra Barchfeld, and Raymond P. Perry. 2011. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology* 36, 1: 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002
170. Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P. Perry. 2002. Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research. *Educational Psychologist* 37, 2: 91–105. https://doi.org/10.1207/S15326985EP3702_4
171. Avar Pentel. 2015. Patterns of Confusion: Using Mouse Logs to Predict User's Emotional State. In *UMAP 2015 Extended Proceedings: Posters, Demos, Late-breaking Results and Workshop Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization (UMAP 2015)*, 40–45. Retrieved January 22, 2017 from http://ceur-ws.org/Vol-1388/PALE2015-paper5.pdf
172. Andrew B. Perry. 2004. Decreasing Math Anxiety in College Students. *College Student Journal* 38, 2: 321.
173. Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10: 1175–1191. https://doi.org/10.1109/34.954607
174. Rosalind Wright Picard. 1995. Affective computing. Retrieved March 7, 2013 from https://www.pervasive.jku.at/Teaching/_2009SS/SeminarausPervasiveComputing/Begleitmaterial/Related%20Work%20(Readings)/1995_Affective%20computing_Picard.pdf
175. Andrea Piscitello. 2015. BiAffect: a System for Analyzing Neurocognitive Functioning Using Keystroke Dynamics and Machine Learning. University of Illinois at Chicago. Retrieved from http://hdl.handle.net/10027/19699
176. Robert Plutchik. 2001. The Nature of Emotions. *American Scientist* 89, 4: 344. https://doi.org/10.1511/2001.4.344
177. Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. 2017. A Survey on Mobile Affective Computing. *Computer Science Review* 25: 79–100. https://doi.org/10.1016/j.cosrev.2017.07.002
178. Kaśka Porayska-Pomsta, Manolis Mavrikis, Sidney D'Mello, Cristina Conati, and Ryan SJd Baker. 2013. Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education* 22, 3: 107–140.

179. Stephen W. Porges. 1997. Emotion: An Evolutionary By-Product of the Neural Regulation of the Autonomic Nervous Systema. *Annals of the New York Academy of Sciences* 807, 1: 62–77. https://doi.org/10.1111/j.1749-6632.1997.tb51913.x
180. Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37, Supplement C: 98–125. https://doi.org/10.1016/j.inffus.2017.02.003
181. Martin F. Porter. 2001. Snowball: A language for stemming algorithms. Retrieved May 10, 2018 from http://snowball.tartarus.org/texts/introduction.html
182. Dave Putwain, Paul Sander, and Derek Larkin. 2013. Academic self-efficacy in study-related skills and behaviours: Relations with learning-related emotions and academic success. *British Journal of Educational Psychology* 83, 4: 633–650. https://doi.org/10.1111/j.2044-8279.2012.02084.x
183. J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
184. Martin Ragot, Nicolas Martin, Sonia Em, Nico Pallamin, and Jean-Marc Diverrez. 2017. Emotion Recognition Using Physiological Signals: Laboratory vs. Wearable Sensors. In *Advances in Human Factors in Wearable Technologies and Game Design* (Advances in Intelligent Systems and Computing), 15–22. https://doi.org/10.1007/978-3-319-60639-2_2
185. Muneeba Raja and Stephan Sigg. 2016. Applicability of RF-based methods for emotion recognition: A survey. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 1–6. https://doi.org/10.1109/PERCOMW.2016.7457119
186. Francesco Riganello, Antonio Candelieri, M. Quintieri, Domenico Conforti, and Giuliano Dolce. 2010. Heart rate variability: An index of brain processing in vegetative state? An artificial intelligence, data mining study. *Clinical Neurophysiology* 121, 12: 2024–2034. https://doi.org/10.1016/j.clinph.2010.05.010
187. Cristóbal Romero, José Raúl Romero, and Sebastián Ventura. 2014. A Survey on Pre-Processing Educational Data. In *Educational Data Mining*, Alejandro Peña-Ayala (ed.). Springer International Publishing, 29–64. Retrieved December 10, 2013 from http://link.springer.com/chapter/10.1007/978-3-319-02738-8_2
188. Cristóbal Romero and Sebastián Ventura. 2017. Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7, 1: e1187. https://doi.org/10.1002/widm.1187
189. James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6: 1161–1178. https://doi.org/10.1037/h0077714
190. Jennifer L. Sabourin, Lucy R. Shores, Bradford W. Mott, and James C. Lester. 2013. Understanding and Predicting Student Self-Regulated Learning Strategies in Game-Based Learning Environments. *International Journal of Artificial Intelligence in Education* 23, 1–4: 94–114. https://doi.org/10.1007/s40593-013-0004-6
191. Sergio Salmeron-Majadas. 2014. Affective standards-based modeling in educational contexts from mining multimodal data sources. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014* (CEUR Workshop Proceedings). Retrieved September 5, 2014 from http://ceur-ws.org/Vol-1181/dc_poster_3.pdf
192. Sergio Salmeron-Majadas, Miguel Arevalillo-Herráez, Olga C. Santos, Mar Saneiro, Raúl Cabestrero, Pilar Quirós, David Arnau, and Jesus G. Boticario. 2015.

Filtering of Spontaneous and Low Intensity Emotions in Educational Contexts. In *Artificial Intelligence in Education*, Cristina Conati, Neil Heffernan, Antonija Mitrovic and M. Felisa Verdejo (eds.). Springer International Publishing, 429–438. Retrieved June 30, 2015 from http://link.springer.com/chapter/10.1007/978-3-319-19773-9_43

193. Sergio Salmeron-Majadas, Ryan S. Baker, Olga C. Santos, and Jesus G. Boticario. 2018. A Machine Learning Approach to Leverage Individual Keyboard and Mouse Interaction Behavior From Multiple Users in Real-World Learning Scenarios. *IEEE Access* 6: 39154–39179. https://doi.org/10.1109/ACCESS.2018.2854966
194. Sergio Salmeron-Majadas, Olga C. Santos, and Jesus G. Boticario. 2013. Affective State Detection in Educational Systems through Mining Multimodal Data Sources. In *6th International Conference on Educational Data Mining*, 348–349. Retrieved February 3, 2014 from http://www.educationaldatamining.org/EDM2013/papers/rn_paper_75.pdf
195. Sergio Salmeron-Majadas, Olga C. Santos, and Jesus G. Boticario. 2013. A Multi-Modal and Data Mining Based Emotion Detection Approach to Build an Affective Open Standards-Based User Model to Support Adaptive Learning Environments. In *Proceedings of the XV Conferencia de La Asociación Española Para La Inteligencia Artificial (CAEPIA)*, 1724–1728.
196. Sergio Salmeron-Majadas, Olga C. Santos, and Jesus G. Boticario. 2013. Inclusive Personalized e-Learning Based on Affective Adaptive Support. In *User Modeling, Adaptation, and Personalization*, Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli and Giovanni Semeraro (eds.). Springer Berlin Heidelberg, 384–387. Retrieved September 5, 2014 from http://link.springer.com/chapter/10.1007/978-3-642-38844-6_45
197. Sergio Salmeron-Majadas, Olga C. Santos, and Jesus G. Boticario. 2014. Exploring indicators from keyboard and mouse interactions to predict the user affective state. In *Proceedings of the 7th International Conference on Educational Data Mining (EDM'14)*, 365–366. Retrieved September 5, 2014 from http://educationaldatamining.org/EDM2014/uploads/procs2014/posters/41_EDM-2014-Poster.pdf
198. Sergio Salmeron-Majadas, Olga C. Santos, and Jesus G. Boticario. 2014. An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context. In *18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, in press.
199. Sergio Salmeron-Majadas, Olga C. Santos, and Jesus G. Boticario. 2014. An Evaluation of Mouse and Keyboard Interaction Indicators towards Non-intrusive and Low Cost Affective Modeling in an Educational Context. *Procedia Computer Science* 35: 691–700. https://doi.org/10.1016/j.procs.2014.08.151
200. Sergio Salmeron-Majadas, Olga C. Santos, and Jesus G. Boticario. 2015. Towards Multimodal Affective Detection in Educational Systems Through Mining Emotional Data Sources. In *Artificial Intelligence in Education*, Cristina Conati, Neil Heffernan, Antonija Mitrovic and M. Felisa Verdejo (eds.). Springer International Publishing, 860–863. Retrieved June 30, 2015 from http://link.springer.com/chapter/10.1007/978-3-319-19773-9_133
201. Sergio Salmeron-Majadas, Olga C. Santos, Jesus G. Boticario, Raúl Cabestrero, Pilar Quirós, and Mar Saneiro. 2013. Gathering Emotional Data from Multiple Sources. In *6th International Conference on Educational Data Mining*, 404–405.

Retrieved February 3, 2014 from http://www.educationaldatamining.org/EDM2013/papers/rn_paper_100.pdf

202. Wendy Sanchez, Alicia Martinez, and Miguel Gonzalez. 2017. Towards Job Stress Recognition Based on Behavior and Physiological Features. In *Ubiquitous Computing and Ambient Intelligence* (Lecture Notes in Computer Science), 311–322. https://doi.org/10.1007/978-3-319-67585-5_33
203. Bonifacio Sandín, Paloma Chorot, Lourdes Lostao, Thomas E. Joiner, Miguel A. Santed, and Rosa M. Valiente. 1999. Escalas PANAS de afecto positivo y negativo: validación factorial y convergencia transcultural. *Psicothema* 11, Número 1: 37–51.
204. Mar Saneiro, Olga C. Santos, Sergio Salmeron-Majadas, and Jesus G. Boticario. 2014. Towards Emotion Detection in Educational Scenarios from Facial Expressions and Body Movements through Multimodal Approaches. *The Scientific World Journal* 2014: e484873. https://doi.org/10.1155/2014/484873
205. Olga C. Santos. 2016. Emotions and Personality in Adaptive e-Learning Systems: An Affective Computing Perspective. In *Emotions and Personality in Personalized Services*. Springer, Cham, 263–285. https://doi.org/10.1007/978-3-319-31413-6_13
206. Olga C. Santos and Jesus G. Boticario. 2014. Exploring Arduino for delivering user contextualized recommendations in e-learning ubiquitous scenarios.
207. Olga C. Santos and Jesus G. Boticario. 2015. Practical guidelines for designing and evaluating educationally oriented recommendations. *Computers & Education* 81: 354–374. https://doi.org/10.1016/j.compedu.2014.10.008
208. Olga C. Santos and Jesus G. Boticario. 2015. Practical guidelines for designing and evaluating educationally oriented recommendations. *Computers & Education* 81, Supplement C: 354–374. https://doi.org/10.1016/j.compedu.2014.10.008
209. Olga C. Santos, Jesus G. Boticario, Miguel Arevalillo-Herráez, Mar Saneiro, Raúl Cabestrero, Elena del Campo, Ángeles Manjarrés, Paloma Moreno-Clari, Pilar Quirós, and Sergio Salmeron-Majadas. 2012. MAMIPEC-Affective Modeling in Inclusive Personalized Educational Scenarios. *Bulletin of the IEEE Technical Committee on Learning Technology* 14, 4: 35.
210. Olga C. Santos, Jesús G. Boticario, Emmanuelle Raffene, and Rafael Pastor. 2007. Why using dotLRN? UNED use cases. In *Proceedings of the FLOSS (Free/Libre/Open Source Systems) International Conference 2007*, 195–212. Retrieved February 28, 2014 from http://www.libros.metabiblioteca.org/bitstream/001/128/8/978-84-9828-124-8.pdf#page=196
211. Olga C. Santos, Alejandro Rodriguez-Ascaso, Jesus G. Boticario, Sergio Salmeron-Majadas, Pilar Quirós, and Raúl Cabestrero. 2013. Challenges for Inclusive Affective Detection in Educational Scenarios. In *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*, Constantine Stephanidis and Margherita Antona (eds.). Springer Berlin Heidelberg, 566–575. Retrieved February 3, 2014 from http://link.springer.com/chapter/10.1007/978-3-642-39188-0_61
212. Olga C. Santos, Sergio Salmeron-Majadas, and Jesus G. Boticario. 2013. Emotions Detection from Math Exercises by Combining Several Data Sources. In *Artificial Intelligence in Education*, H. Chad Lane, Kalina Yacef, Jack Mostow and Philip Pavlik (eds.). Springer Berlin Heidelberg, 742–745. Retrieved October 4, 2013 from http://link.springer.com/chapter/10.1007/978-3-642-39112-5_102

213. Olga C. Santos, Mar Saneiro, Jesus G. Boticario, and M. C. Rodriguez-Sanchez. 2016. Toward interactive context-aware affective educational recommendations in computer-assisted language learning. *New Review of Hypermedia and Multimedia* 22, 1–2: 27–57. https://doi.org/10.1080/13614568.2015.1058428
214. Olga C. Santos, Mar Saneiro, Sergio Salmeron-Majadas, and Jesus G. Boticario. 2014. A Methodological Approach to Eliciting Affective Educational Recommendations. In *2014 IEEE 14th International Conference on Advanced Learning Technologies (ICALT)*, 529–533. https://doi.org/10.1109/ICALT.2014.234
215. Olga C. Santos, Raul Uria-Rivas, Cristina Rodriguez-Sanchez, and Jesus G. Boticario. 2016. An Open Sensing and Acting Platform for Context-Aware Affective Support in Ambient Intelligent Educational Settings. *IEEE Sensors Journal* 16, 10: 3865–3874. https://doi.org/10.1109/JSEN.2016.2533266
216. Martin Schätz, Fabio Centonze, Jirí Kuchyňka, Ondrej Ťupa, Oldrich Vyšata, Oana Geman, and Ales Procházka. 2015. Statistical recognition of breathing by MS Kinect depth sensor. In *2015 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*, 1–4. https://doi.org/10.1109/IWCIM.2015.7347062
217. Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and Metrics for Cold-start Recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '02), 253–260. https://doi.org/10.1145/564376.564421
218. Harold Schlosberg. 1952. The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology* 44, 4: 229–237. https://doi.org/10.1037/h0055778
219. Harold Schlosberg. 1954. Three dimensions of emotion. *Psychological Review* 61, 2: 81–88. https://doi.org/10.1037/h0054570
220. Marc Schröder, Paolo Baggia, Felix Burkhardt, Catherine Pelachaud, Christian Peter, and Enrico Zovato. 2011. EmotionML – An Upcoming Standard for Representing Emotions and Related States. In *Affective Computing and Intelligent Interaction*, Sidney D'Mello, Arthur Graesser, Björn Schuller and Jean-Claude Martin (eds.). Springer Berlin Heidelberg, 316–325. Retrieved September 2, 2014 from http://link.springer.com/chapter/10.1007/978-3-642-24600-5_35
221. Marc Schröeder, Paolo Baggia, Felix Burkhardt, Catherine Pelachaud, Christian Peter, and Enrico Zovato. 2010. Emotion markup language (EmotionML) 1.0. *W3C Working Draft* 29: 3–22.
222. Ralf Schwarzer. 1993. *Measurement of perceived self-efficacy: Psychometric scales for cross-cultural research*. Freien Universität.
223. Ralf Schwarzer and Judith Baessler. 1996. Evaluación de la autoeficacia: adaptación española de la Escala de Autoeficacia general. *Ansiedad y estrés* 2, 1: 1–8.
224. S. Seneviratne, Y. Hu, T. Nguyen, G. Lan, S. Khalifa, K. Thilakarathna, M. Hassan, and A. Seneviratne. 2017. A Survey of Wearable Devices and Challenges. *IEEE Communications Surveys Tutorials* 19, 4: 2573–2620. https://doi.org/10.1109/COMST.2017.2731979
225. Sergio Salmeron-Majadas. 2014. Addressing the Difficulty of Emotions Detection in Educational Scenarios from a Multimodal Data Mining Approach. UNED, UNED University, Madrid.

226. David Sheffield and Thomas Hunt. 2006. How does Anxiety Influence Maths Performance and what can we do about it? *MSOR Connections* 6, 4: 19–23.
227. Rayhan Shikder, Sydur Rahaman, Farzia Afroze, and A. B. M. Alim Al Islam. 2017. Keystroke/mouse usage based emotion detection and user identification. In *2017 International Conference on Networking, Systems and Security (NSysS)*, 96–104. https://doi.org/10.1109/NSysS.2017.7885808
228. Pragya Shukla and Rinky Solanki. 2013. Web Based Keystroke Dynamics Application for Identifying Emotional State. *International Journal of Advanced Research in Computer and Communication Engineering* 2, 11: 4489–4493.
229. Paul J. Silvia. 2002. Self-awareness and emotional intensity. *Cognition & Emotion* 16, 2: 195–216. https://doi.org/10.1080/02699930143000310
230. Robert A. Sottilare and Michael Proctor. 2012. Passively Classifying Student Mood and Performance within Intelligent Tutors. *Educational Technology & Society* 15, 2: 101–114.
231. P. D. Sturkie. 1986. Heart: Contraction, Conduction, and Electrocardiography. In *Avian Physiology*. Springer, New York, NY, 167–190. https://doi.org/10.1007/978-1-4612-4862-0_7
232. Melanie Swan. 2013. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data* 1, 2: 85–99. https://doi.org/10.1089/big.2012.0002
233. Juan Carlos Cobos Torres and Mohamed Abderrahim. 2017. Measuring Heart and Breath Rates by Image Photoplethysmography using Wavelets Technique. *IEEE Latin America Transactions* 15, 10: 1864–1868. https://doi.org/10.1109/TLA.2017.8071228
234. Radek Trnka, Karel Balcar, Martin Kuska, and Czech Science Foundation. 2011. *Re-constructing emotional spaces from experience to regulation*. Prague College of Psychosocial Studies Press, Prague.
235. Karen M. Trujillo and Oakley D. Hadfield. 1999. Tracing the Roots of Mathematics Anxiety through In-Depth Interviews with Preservice Elementary Teachers. *College Student Journal* 33, 2: 219.
236. Georgios Tsoulouhas, Dimitrios Georgiou, and Alexandros Karakos. 2011. Detection of Learner’s Affective State Based on Mouse Movements. *Journal of Computing* 3, 11: 9–18.
237. Wei-Hsuan Tsui, Poming Lee, and Tzu-Chien Hsiao. 2013. The effect of emotion on keystroke: An experimental study using facial feedback hypothesis. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2870–2873. https://doi.org/10.1109/EMBC.2013.6610139
238. Mindaugas Ulinskas, Marcin Woźniak, and Robertas Damaševičius. 2017. Analysis of Keystroke Dynamics for Fatigue Recognition. In *Computational Science and Its Applications – ICCSA 2017* (Lecture Notes in Computer Science), 235–247. https://doi.org/10.1007/978-3-319-62404-4_18
239. Larry Vea and Ma Mercedes Rodrigo. 2016. Modeling Negative Affect Detector of Novice Programming Students Using Keyboard Dynamics and Mouse Behavior. In *Trends in Artificial Intelligence: PRICAI 2016 Workshops* (Lecture Notes in Computer Science), 127–138. https://doi.org/10.1007/978-3-319-60675-0_11
240. María Villarejo, Begoña Zapirain, and Amaia Zorrilla. 2013. Algorithms Based on CWT and Classifiers to Control Cardiac Alterations and Stress Using an ECG and a SCR. *Sensors* 13, 5: 6141–6170. https://doi.org/10.3390/s130506141

241. Lisa M. Vizer, Lina Zhou, and Andrew Sears. 2009. Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies* 67, 10: 870–886. https://doi.org/10.1016/j.ijhcs.2009.07.005
242. Y. Wang, T. Wang, P. Gong, Y. Wu, C. Ye, J. Li, and T. Ma. 2017. A multi-label learning method for efficient affective detection. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 61–64. https://doi.org/10.1109/BHI.2017.7897205
243. David Watson, Lee A. Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6: 1063.
244. Joseph B. Wiggins, Joseph F. Grafsgaard, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. 2014. The Relationship Between Task Difficulty and Emotion in Online Computer Programming Tutoring. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education* (SIGCSE ’14), 721–721. https://doi.org/10.1145/2538862.2544298
245. Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2, 1–3: 37–52. https://doi.org/10.1016/0169-7439(87)80084-9
246. Wu Chih-Hung, Huang Yueh-Min, and Hwang Jan-Pan. 2015. Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology* 47, 6: 1304–1323. https://doi.org/10.1111/bjet.12324
247. Wilhelm Max Wundt. 1897. *Outlines of psychology*. Leipzig, W. Engelmann; New York, G.E. Stechert. Retrieved September 5, 2014 from http://archive.org/details/cu31924014474534
248. Takashi Yamauchi. 2013. Mouse Trajectories and State Anxiety: Feature Selection with Random Forest. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 399–404. https://doi.org/10.1109/ACII.2013.72
249. Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. 2017. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer Methods and Programs in Biomedicine* 140: 93–110. https://doi.org/10.1016/j.cmpb.2016.12.005
250. Li Yuan, Stephen Powell, and JISC CETIS. 2013. MOOCs and open education: Implications for higher education. *Cetis White Paper*. Retrieved September 1, 2014 from http://www.smarthighered.com/wp-content/uploads/2013/03/MOOCs-and-Open-Education.pdf
251. Jing Zhai and Armando Barreto. 2006. Stress Recognition Using Non-invasive Technology. In *FLAIRS Conference*, 395–401. Retrieved November 20, 2013 from http://www.aaai.org/Papers/FLAIRS/2006/Flairs06-077.pdf
252. Shengkai Zhang and Pan Hui. 2014. A Survey on Mobile Affective Computing. *arXiv:1410.1648 [cs]* 1410: arXiv:1410.1648.
253. Yu Zhong and Yunbin Deng. 2015. A survey on keystroke dynamics biometrics: approaches, advances, and evaluations. *Recent Advances in User Authentication Using Keystroke Dynamics Biometrics. Science Gate Publishing*: 1–22.
254. Philippe Zimmermann, Patrick Gomez, Brigitta Danuser, and S. Schär. 2006. Extending usability: putting affect into the user-experience. *Proceedings of NordiCHI’06*: 27–32.
255. Philippe Zimmermann, Sissel Guttormsen, Brigitta Danuser, and Patrick Gomez. 2003. Affective computing–a rationale for measuring mood with mouse and

keyboard. *International journal of occupational safety and ergonomics* 9, 4: 539–551.

256. 2014. UNED: Nuestra Historia. *UNED: Nuestra Historia*. Retrieved September 1, 2014 from http://portal.uned.es/portal/page?_pageid=93,499271&_dad=portal&_schema=PORTAL
257. Schröder M. and Pelachaud C., Vocabularies for EmotionML. Editors. W3C Working Group Note, 10 May 2012.

13. Appendices

13.1. Appendix I: Full list of published works from this research

Here the references of all the published works including material related to the research are to be presented.

13.1.1. Publications driven by the author of the work here presented

- **Sergio Salmeron-Majadas**, Ryan S. Baker, Olga C. Santos, and Jesus G. Boticario. 2018. A Machine Learning Approach to Leverage Individual Keyboard and Mouse Interaction Behavior from Multiple Users in Real-World Learning Scenarios. IEEE Access (2018). DOI:https://doi.org/10.1109/ACCESS.2018.2854966
- **Sergio Salmeron-Majadas**, Miguel Arevalillo-Herráez, Olga C. Santos, Mar Saneiro, Raúl Cabestrero, Pilar Quirós, David Arnau, and Jesus G. Boticario. 2015. Filtering of Spontaneous and Low Intensity Emotions in Educational Contexts. In Artificial Intelligence in Education, Cristina Conati, Neil Heffernan, Antonija Mitrovic and M. Felisa Verdejo (eds.). Springer International Publishing, 429–438. Retrieved June 30, 2015 from http://link.springer.com/chapter/10.1007/978-3-319-19773-9_43
- **Sergio Salmeron-Majadas**, Olga C. Santos, and Jesus G. Boticario. 2015. Towards Multimodal Affective Detection in Educational Systems Through Mining Emotional Data Sources. In Artificial Intelligence in Education, Cristina Conati, Neil Heffernan, Antonija Mitrovic and M. Felisa Verdejo (eds.). Springer International Publishing, 860–863. Retrieved June 30, 2015 from http://link.springer.com/chapter/10.1007/978-3-319-19773-9_133
- **Sergio Salmeron-Majadas**. 2014. Affective standards-based modeling in educational contexts from mining multimodal data sources. In Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization co-located with the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP2014), Aalborg, Denmark, July 7-11, 2014 (CEUR Workshop Proceedings). Retrieved September 5, 2014 from http://ceur-ws.org/Vol-1181/dc_poster_3.pdf
- **Sergio Salmeron-Majadas**, Olga C. Santos, and Jesus G. Boticario. 2014. Exploring indicators from keyboard and mouse interactions to predict the user affective state. In Proceedings of the 7th International Conference on Educational Data Mining (EDM'14), 365–366. Retrieved September 5, 2014

from http://educationaldatamining.org/EDM2014/uploads/procs2014/posters/41_EDM-2014-Poster.pdf

- **Sergio Salmeron-Majadas**, Olga C. Santos, and Jesus G. Boticario. 2014. An Evaluation of Mouse and Keyboard Interaction Indicators towards Non-intrusive and Low Cost Affective Modeling in an Educational Context. Procedia Computer Science 35, (January 2014), 691–700. DOI:https://doi.org/10.1016/j.procs.2014.08.151
- **Sergio Salmeron-Majadas**, Olga C. Santos, Jesus G. Boticario, Raúl Cabestrero, Pilar Quirós, and Mar Saneiro. 2013. Gathering Emotional Data from Multiple Sources. In 6th International Conference on Educational Data Mining, 404–405. Retrieved February 3, 2014 from http://www.educationaldatamining.org/EDM2013/papers/rn_paper_100.pdf
- **Sergio Salmeron-Majadas**, Olga C. Santos, and Jesus G. Boticario. 2013. Affective State Detection in Educational Systems through Mining Multimodal Data Sources. In 6th International Conference on Educational Data Mining, 348–349. Retrieved February 3, 2014 from http://www.educationaldatamining.org/EDM2013/papers/rn_paper_75.pdf
- **Sergio Salmeron-Majadas**, Olga C. Santos, and Jesus G. Boticario. 2013. A Multi-Modal and Data Mining Based Emotion Detection Approach to Build an Affective Open Standards-Based User Model to Support Adaptive Learning Environments. In Proceedings of the XV Conferencia de La Asociación Española Para La Inteligencia Artificial (CAEPIA), 1724–1728.
- **Sergio Salmeron-Majadas**, Olga C. Santos, and Jesus G. Boticario. 2013. Inclusive Personalized e-Learning Based on Affective Adaptive Support. In User Modeling, Adaptation, and Personalization, Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli and Giovanni Semeraro (eds.). Springer Berlin Heidelberg, 384–387. Retrieved November 25, 2014 from http://link.springer.com/chapter/10.1007/978-3-642-38844-6_45

13.1.2. Related publications where the author of the work here presented collaborated

- Raúl Cabestrero, Pilar Quirós, Olga C. Santos, **Sergio Salmeron-Majadas**, Raul Uria-Rivas, Jesus G. Boticario, David Arnau, Miguel Arevalillo-Herráez, and Francesc J. Ferri. 2018. Some insights into the impact of affective information when delivering feedback to students. Behaviour & Information Technology 0, 0 (July 2018), 1–12. DOI:https://doi.org/10.1080/0144929X.2018.1499803
- Jesus G. Boticario, Olga C. Santos, Raúl Cabestrero, Pilar Quirós, **Sergio Salmerón-Majadas,** Raúl Uria-Rivas, Mar Saneiro, Miguel Arevalillo-Herráez, and Francesc J. Ferri. 2017. BIG-AFF: Exploring Low Cost and Low Intrusive Infrastructures for Affective Computing in Secondary Schools. In Adjunct Publication of the 25th Conference on User Modeling, Adaptation and

Personalization, 287–292. Retrieved from http://dl.acm.org/citation.cfm?id=3099084

- Luis Marco-Giménez, Miguel Arevalillo-Herráez, Francesc J. Ferri, Salvador Moreno-Picot, Jesus Boticario, Olga C. Santos, **Sergio Salmeron-Majadas**, Mar Saneiro, Raul Uria-Rivas, David Arnau, and others. 2016. Affective and Behavioral Assessment for Adaptive Intelligent Tutoring Systems. In UMAP (Extended Proceedings). Retrieved from http://ceur-ws.org/Vol-1618/PALE3.pdf
- Luis Marco-Gimenez, Migueñ Arevalillo-Herraez, Francesc J. Ferri, Salvador Moreno-Picot, Jesus G. Boticario, Olga C. Santos, **Sergio Salmeron-Majadas**, Mar Saneiro, Raul Uria-Rivas, David Arnau, and others. 2015. Affective and Behavioral Assessment for Adaptive Intelligent Tutoring Systems. Personalization Approaches in Learning Environments (2015). Retrieved July 13, 2016 from http://ceur-ws.org/Vol-1618/PALE3.pdf
- Olga C. Santos, Mar Saneiro, M. Cristina Rodriguez-Sanchez, Jesus G. Boticario, Raul Uria-Rivas, and **Sergio Salmeron-Majadas**. 2015. The potential of Ambient Intelligence to deliver Interactive Context-Aware Affective Educational support through Recommendations. In International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2015), 1. Retrieved May 24, 2016 from http://ceur-ws.org/Vol-1446/amadl_pap1.pdf
- Miguel Arevalillo-Herráez, David Arnau, Jesus G. Boticario, José Antonio González-Calero, Paloma Moreno-Clari, Salvador Moreno-Picot, **Sergio Salmeron-Majadas**, and Olga C. Santos. 2014. Computación afectiva: desarrollos y propuestas de uso en la construcción de sistemas de tutorización inteligente.
- Miguel Arevalillo-Herráez, David Arnau, Luis Marco-Giménez, José A. González-Calero, Salvador Moreno-Picot, Paloma Moreno-Clari, Aladdin Ayesh, Olga C. Santos, Jesús G. Boticario, Mar Saneiro, **Sergio Salmeron-Majadas**, Pilar Quirós, and Raúl Cabestrero. 2014. Providing Personalized Guidance in Arithmetic Problem Solving. In Personalization Approaches in Learning Environments, 42–48. Retrieved September 5, 2014 from http://ceur-ws.org/Vol-1181/pale2014_paper_05.pdf
- Jesus G. Boticario, Olga C. Santos, Raúl Cabestrero, Pilar Quirós, Mar Saneiro, **Sergio Salmeron-Majadas**, Ángeles Manjarrés, Alejandro Rodriguez-Ascaso, Elena del Campo, and Emmanuelle Raffene. 2014. Avances en el modelado de aspectos afectivos en escenarios educativos inclusivos y personalizados. (2014).
- Mar Saneiro, Olga C. Santos, **Sergio Salmeron-Majadas**, and Jesus G. Boticario. 2014. Towards Emotion Detection in Educational Scenarios from Facial Expressions and Body Movements through Multimodal Approaches. The Scientific World Journal 2014, (April 2014), e484873. DOI:https://doi.org/10.1155/2014/484873
- Olga C. Santos, Mar Saneiro, **Sergio Salmeron-Majadas**, and Jesus G. Boticario. 2014. A Methodological Approach to Eliciting Affective Educational

Recommendations. In 2014 IEEE 14th International Conference on Advanced Learning Technologies (ICALT), 529–533. DOI:https://doi.org/10.1109/ICALT.2014.234

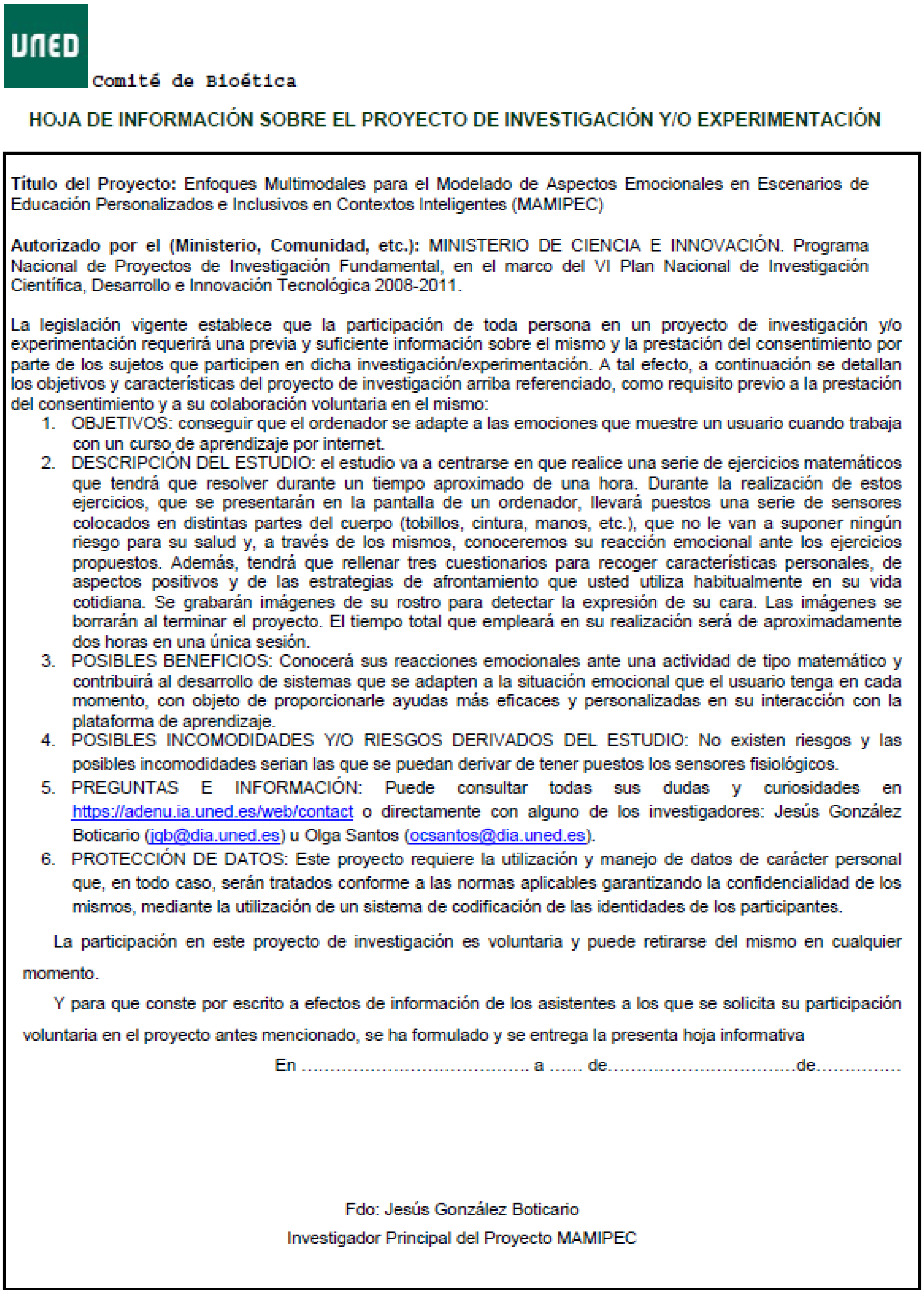
- Olga C. Santos, **Sergio Salmeron-Majadas**, and Jesus G. Boticario. 2014. Supporting Growers with Recommendations in RedVides: Some Human Aspects Involved. In Web-Age Information Management, Yueguo Chen, Wolf-Tilo Balke, Jianliang Xu, Wei Xu, Peiquan Jin, Xin Lin, Tiffany Tang and Eenjun Hwang (eds.). Springer International Publishing, 307–314. Retrieved April 20, 2015 from http://link.springer.com/chapter/10.1007/978-3-319-11538-2_28
- Miguel Arevalillo-herráez, Salvador Moreno-picot, David Arnau, Paloma Moreno, Jesus G. Boticario, Olga C. Santos, Raúl Cabestrero, Pilar Quirós, **Sergio Salmeron-Majadas**, Ángeles Manjarrés-riesco, and Mar Saneiro. 2013. Towards Enriching an ITS with Affective Support. In Personalization Approaches in Learning Environments, 5–12. Retrieved from http://ceur-ws.org/Vol-997/pale2013_paper_1.pdf
- Olga C. Santos, Alejandro Rodriguez-Ascaso, Jesus G. Boticario, **Sergio Salmeron-Majadas**, Pilar Quirós, and Raúl Cabestrero. 2013. Challenges for Inclusive Affective Detection in Educational Scenarios. In Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion, Constantine Stephanidis and Margherita Antona (eds.). Springer Berlin Heidelberg, 566–575. Retrieved November 25, 2014 from http://link.springer.com/chapter/10.1007/978-3-642-39188-0_61
- Olga C. Santos, **Sergio Salmeron-Majadas**, and Jesus G. Boticario. 2013. Emotions Detection from Math Exercises by Combining Several Data Sources. In Artificial Intelligence in Education, H. Chad Lane, Kalina Yacef, Jack Mostow and Philip Pavlik (eds.). Springer Berlin Heidelberg, 742–745. Retrieved April 2, 2014 from http://link.springer.com/chapter/10.1007/978-3-642-39112-5_102
- Olga C. Santos, Mar Saneiro, Emmanuelle Gutiérrez y Restrepo, Jesus Boticario, Elena Del Campo, Raúl Cabestrero, Pilar Quirós, **Sergio Salmeron-Majadas**, Emmanuelle Raffenne, and Emanuela Mazzone. 2013. techplay. mobi: A Technological Framework for Developing Affective Inclusive Personalized Mobile Serious Games to Enrich Learning Competences. In Extended Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization. Retrieved November 8, 2016 from https://pdfs.semanticscholar.org/e858/a27fee4efef5ebe500b485ae82dc057d8442.pdf
- Olga C. Santos, Jesus G. Boticario, Miguel Arevalillo-Herráez, Mar Saneiro, Raúl Cabestrero, Elena del Campo, Ángeles Manjarrés, Paloma Moreno-Clari, Pilar Quirós, and **Sergio Salmeron-Majadas**. 2012. MAMIPEC-Affective Modeling in Inclusive Personalized Educational Scenarios. Bulletin of the IEEE Technical Committee on Learning Technology 14, 4 (2012), 35.

13.2. **Appendix II : Experimental material**

This section compiles the material use in the experiments carried out. As the experiment was addressed to citizens from Madrid (Spain), contents were provided in Spanish. The following materials are attached:

- Information consent
- Demographic and psychological questionnaires
- Calibration questions
- Calibration images
- Calibration sounds
- Math exercises used in stage 1 experiment
- Graphical logical series
- Satisfaction questionnaire and PANAS

13.2.1. Information Consent



UNED
Comité de Bioética

HOJA DE INFORMACIÓN SOBRE EL PROYECTO DE INVESTIGACIÓN Y/O EXPERIMENTACIÓN

Título del Proyecto: Enfoques Multimodales para el Modelado de Aspectos Emocionales en Escenarios de Educación Personalizados e Inclusivos en Contextos Inteligentes (MAMIPEC)

Autorizado por el (Ministerio, Comunidad, etc.): MINISTERIO DE CIENCIA E INNOVACIÓN. Programa Nacional de Proyectos de Investigación Fundamental, en el marco del VI Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008-2011.

La legislación vigente establece que la participación de toda persona en un proyecto de investigación y/o experimentación requerirá una previa y suficiente información sobre el mismo y la prestación del consentimiento por parte de los sujetos que participen en dicha investigación/experimentación. A tal efecto, a continuación se detallan los objetivos y características del proyecto de investigación arriba referenciado, como requisito previo a la prestación del consentimiento y a su colaboración voluntaria en el mismo:

1. OBJETIVOS: conseguir que el ordenador se adapte a las emociones que muestre un usuario cuando trabaja con un curso de aprendizaje por internet.
2. DESCRIPCIÓN DEL ESTUDIO: el estudio va a centrarse en que realice una serie de ejercicios matemáticos que tendrá que resolver durante un tiempo aproximado de una hora. Durante la realización de estos ejercicios, que se presentarán en la pantalla de un ordenador, llevará puestos una serie de sensores colocados en distintas partes del cuerpo (tobillos, cintura, manos, etc.), que no le van a suponer ningún riesgo para su salud y, a través de los mismos, conoceremos su reacción emocional ante los ejercicios propuestos. Además, tendrá que rellenar tres cuestionarios para recoger características personales, de aspectos positivos y de las estrategias de afrontamiento que usted utiliza habitualmente en su vida cotidiana. Se grabarán imágenes de su rostro para detectar la expresión de su cara. Las imágenes se borrarán al terminar el proyecto. El tiempo total que empleará en su realización será de aproximadamente dos horas en una única sesión.
3. POSIBLES BENEFICIOS: Conocerá sus reacciones emocionales ante una actividad de tipo matemático y contribuirá al desarrollo de sistemas que se adapten a la situación emocional que el usuario tenga en cada momento, con objeto de proporcionarle ayudas más eficaces y personalizadas en su interacción con la plataforma de aprendizaje.
4. POSIBLES INCOMODIDADES Y/O RIESGOS DERIVADOS DEL ESTUDIO: No existen riesgos y las posibles incomodidades serian las que se puedan derivar de tener puestos los sensores fisiológicos.
5. PREGUNTAS E INFORMACIÓN: Puede consultar todas sus dudas y curiosidades en https://adenu.ia.uned.es/web/contact o directamente con alguno de los investigadores: Jesús González Boticario (jgb@dia.uned.es) u Olga Santos (ocsantos@dia.uned.es).
6. PROTECCIÓN DE DATOS: Este proyecto requiere la utilización y manejo de datos de carácter personal que, en todo caso, serán tratados conforme a las normas aplicables garantizando la confidencialidad de los mismos, mediante la utilización de un sistema de codificación de las identidades de los participantes.

La participación en este proyecto de investigación es voluntaria y puede retirarse del mismo en cualquier momento.

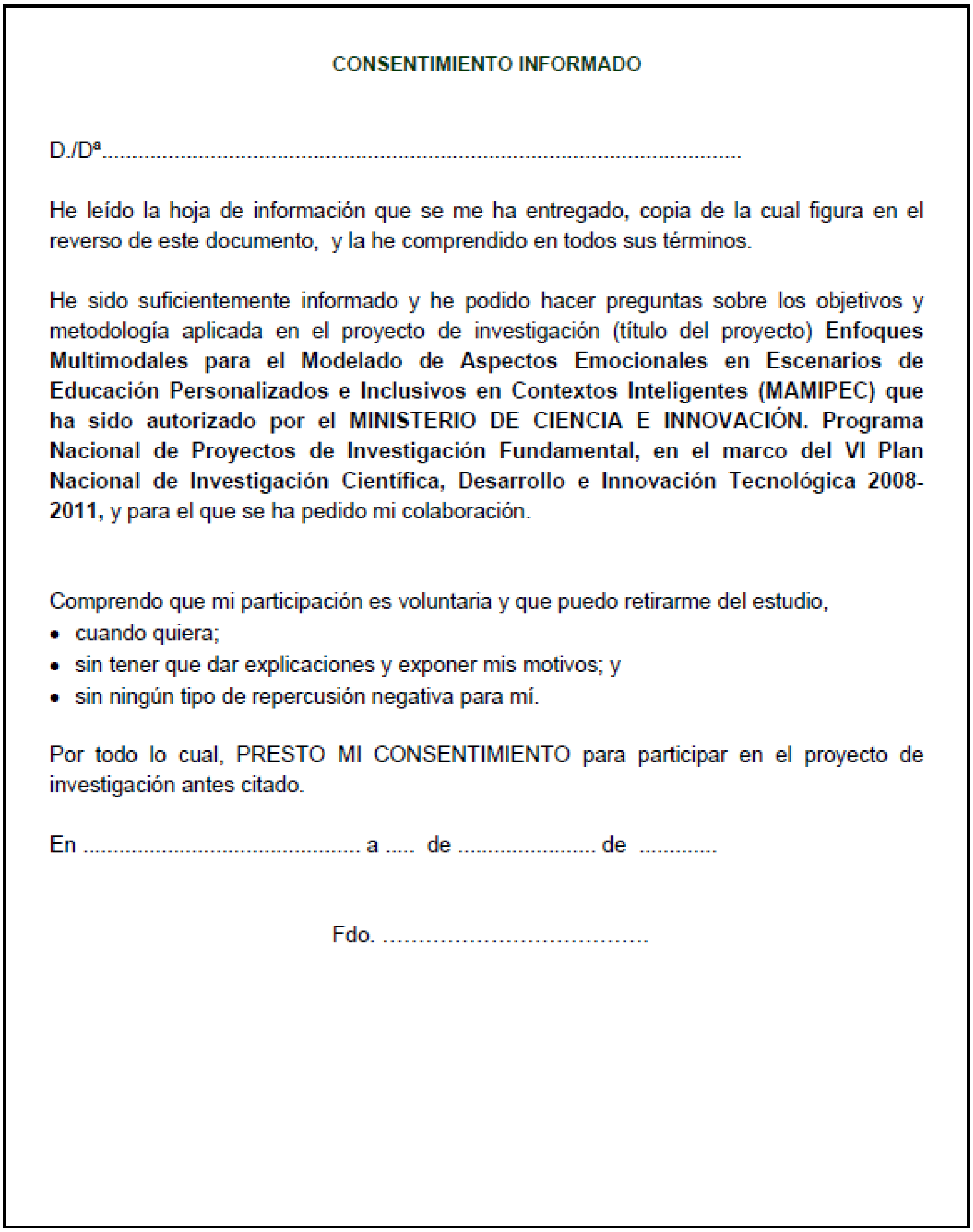
Y para que conste por escrito a efectos de información de los asistentes a los que se solicita su participación voluntaria en el proyecto antes mencionado, se ha formulado y se entrega la presenta hoja informativa

En .... a ...... de................................de...............

Fdo: Jesús González Boticario

Investigador Principal del Proyecto MAMIPEC

Figure 69. Information consent (page 1).



CONSENTIMIENTO INFORMADO

D./Dª...

He leído la hoja de información que se me ha entregado, copia de la cual figura en el reverso de este documento, y la he comprendido en todos sus términos.

He sido suficientemente informado y he podido hacer preguntas sobre los objetivos y metodología aplicada en el proyecto de investigación (título del proyecto) **Enfoques Multimodales para el Modelado de Aspectos Emocionales en Escenarios de Educación Personalizados e Inclusivos en Contextos Inteligentes (MAMIPEC) que ha sido autorizado por el MINISTERIO DE CIENCIA E INNOVACIÓN. Programa Nacional de Proyectos de Investigación Fundamental, en el marco del VI Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008-2011,** y para el que se ha pedido mi colaboración.

Comprendo que mi participación es voluntaria y que puedo retirarme del estudio,

- cuando quiera;
- sin tener que dar explicaciones y exponer mis motivos; y
- sin ningún tipo de repercusión negativa para mí.

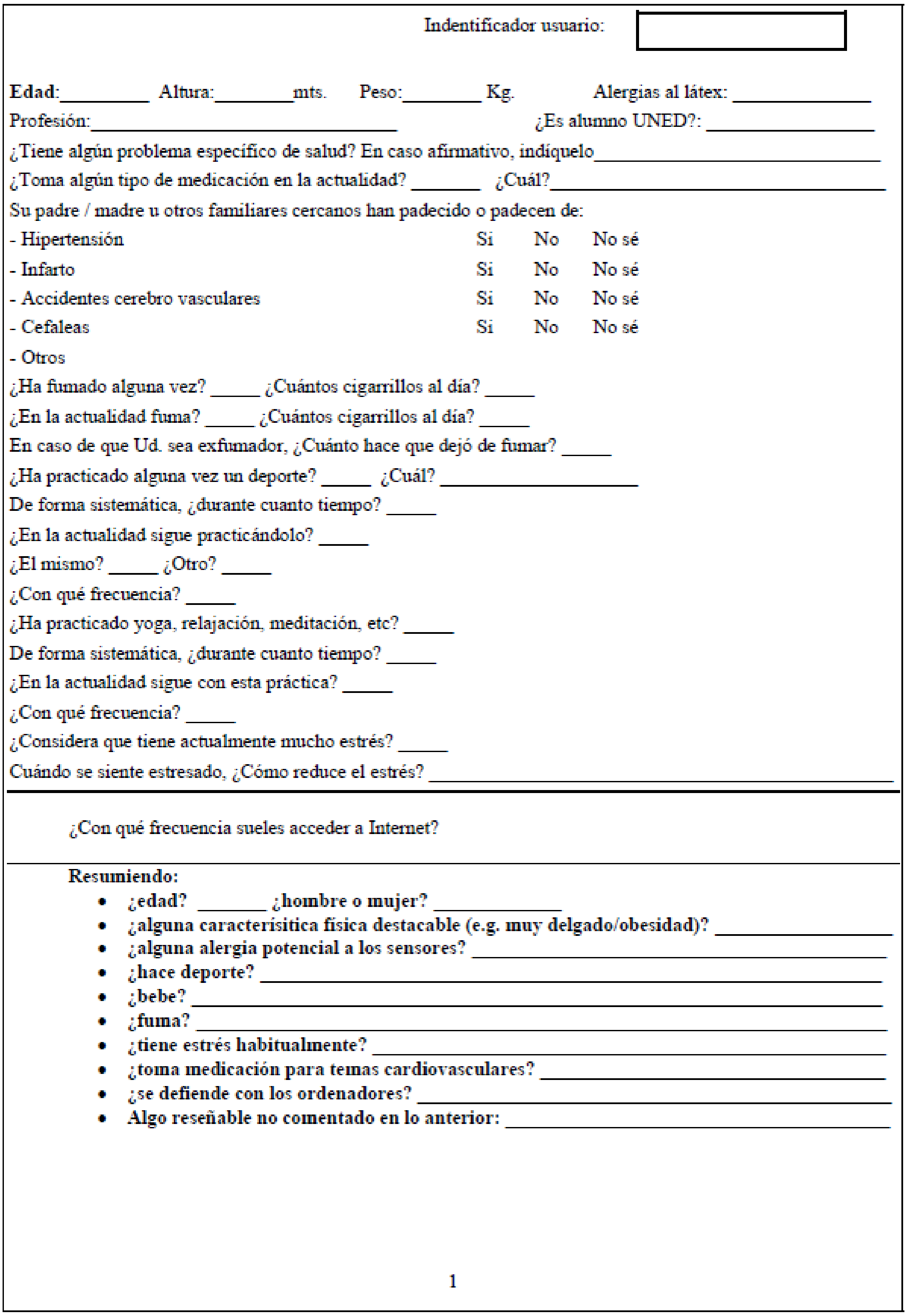
Por todo lo cual, PRESTO MI CONSENTIMIENTO para participar en el proyecto de investigación antes citado.

En .... a ..... de .... de ....

Fdo. ....

Figure 70.Information consent (page 2).

13.2.2. Demographic questionnaire



Indentificador usuario: □

Edad:_________ Altura:________mts. Peso:________ Kg. Alergias al látex: ______________

Profesión:________________________________ ¿Es alumno UNED?: _________________

¿Tiene algún problema específico de salud? En caso afirmativo, indíquelo_____________________________

¿Toma algún tipo de medicación en la actualidad? _______ ¿Cuál?__________________________________

Su padre / madre u otros familiares cercanos han padecido o padecen de:

- Hipertensión Si No No sé
- Infarto Si No No sé
- Accidentes cerebro vasculares Si No No sé
- Cefaleas Si No No sé
- Otros

¿Ha fumado alguna vez? _____ ¿Cuántos cigarrillos al día? _____

¿En la actualidad fuma? _____ ¿Cuántos cigarrillos al día? _____

En caso de que Ud. sea exfumador, ¿Cuánto hace que dejó de fumar? _____

¿Ha practicado alguna vez un deporte? _____ ¿Cuál? ____________________

De forma sistemática, ¿durante cuanto tiempo? _____

¿En la actualidad sigue practicándolo? _____

¿El mismo? _____ ¿Otro? _____

¿Con qué frecuencia? _____

¿Ha practicado yoga, relajación, meditación, etc? _____

De forma sistemática, ¿durante cuanto tiempo? _____

¿En la actualidad sigue con esta práctica? _____

¿Con qué frecuencia? _____

¿Considera que tiene actualmente mucho estrés? _____

Cuándo se siente estresado, ¿Cómo reduce el estrés? ____

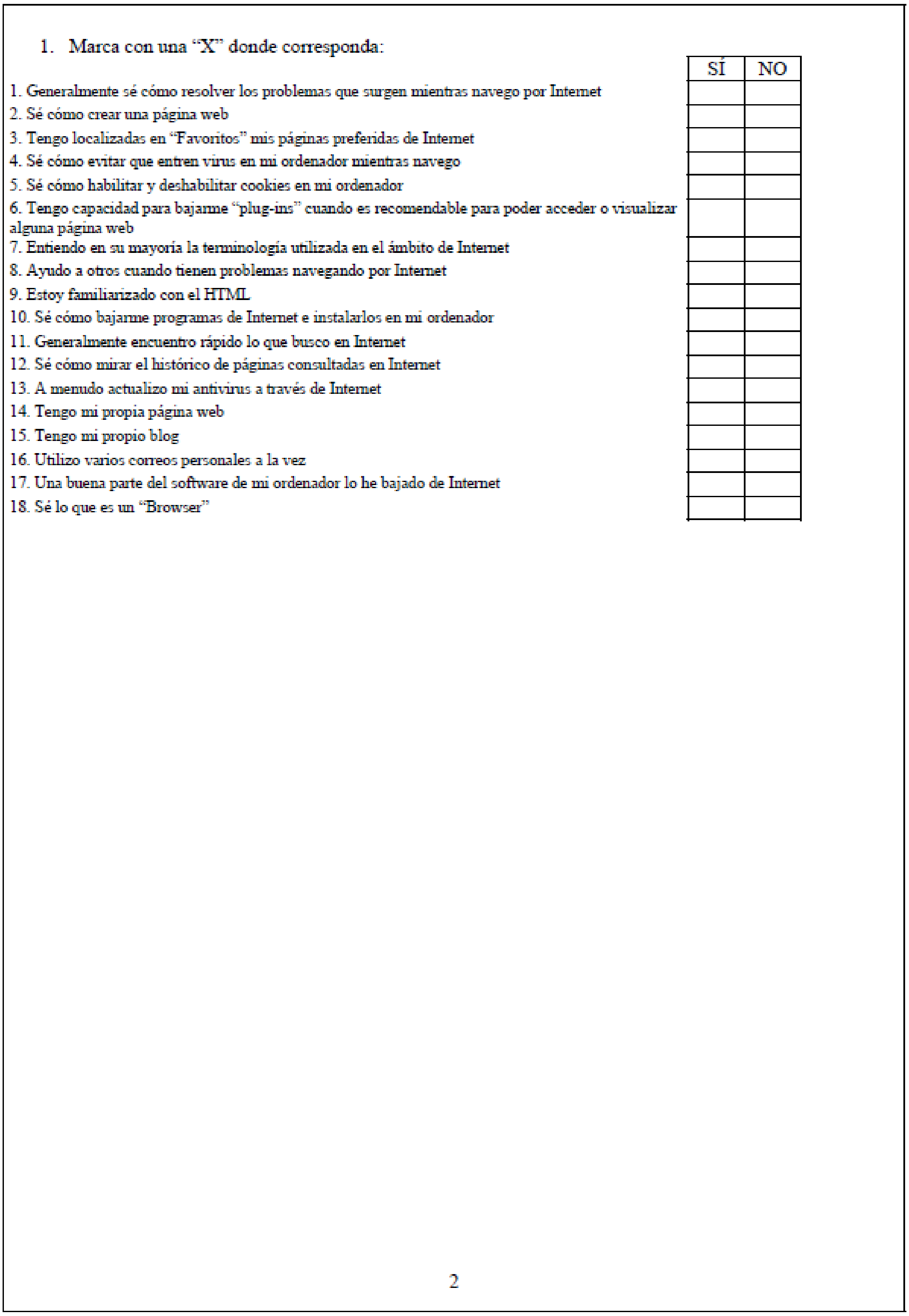
¿Con qué frecuencia sueles acceder a Internet?

Resumiendo:

- **¿edad? _______ ¿hombre o mujer? _____________**
- **¿alguna caracterísitica física destacable (e.g. muy delgado/obesidad)? __________________**
- **¿alguna alergia potencial a los sensores? _______________________________________**
- **¿hace deporte? ____**
- **¿bebe? ____**
- **¿fuma? ____**
- **¿tiene estrés habitualmente? ____**
- **¿toma medicación para temas cardiovasculares? ___________________________________**
- **¿se defiende con los ordenadores? ____**
- **Algo reseñable no comentado en lo anterior: _______________________________________**

1

Figure 71. Demographic Questionnaire (page 1).



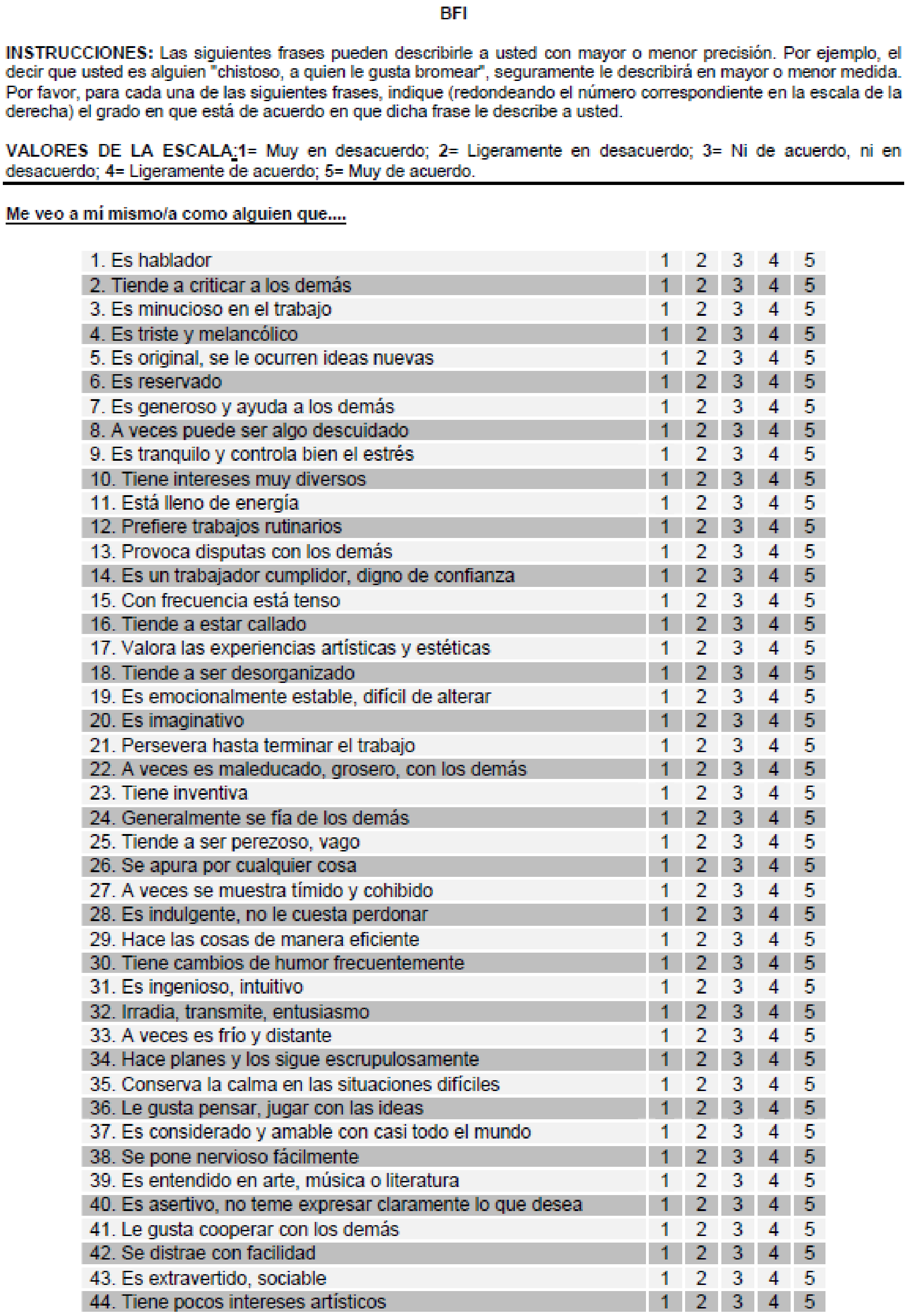
1. Marca con una “X” donde corresponda:

	SÍ	NO
1. Generalmente sé cómo resolver los problemas que surgen mientras navego por Internet		
2. Sé cómo crear una página web		
3. Tengo localizadas en “Favoritos” mis páginas preferidas de Internet		
4. Sé cómo evitar que entren virus en mi ordenador mientras navego		
5. Sé cómo habilitar y deshabilitar cookies en mi ordenador		
6. Tengo capacidad para bajarme “plug-ins” cuando es recomendable para poder acceder o visualizar alguna página web		
7. Entiendo en su mayoría la terminología utilizada en el ámbito de Internet		
8. Ayudo a otros cuando tienen problemas navegando por Internet		
9. Estoy familiarizado con el HTML		
10. Sé cómo bajarme programas de Internet e instalarlos en mi ordenador		
11. Generalmente encuentro rápido lo que busco en Internet		
12. Sé cómo mirar el histórico de páginas consultadas en Internet		
13. A menudo actualizo mi antivirus a través de Internet		
14. Tengo mi propia página web		
15. Tengo mi propio blog		
16. Utilizo varios correos personales a la vez		
17. Una buena parte del software de mi ordenador lo he bajado de Internet		
18. Sé lo que es un “Browser”		

2

Figure 72. Demographic Questionnaire (page 2).

13.2.3. Personality traits: BFI & GSE



BFI

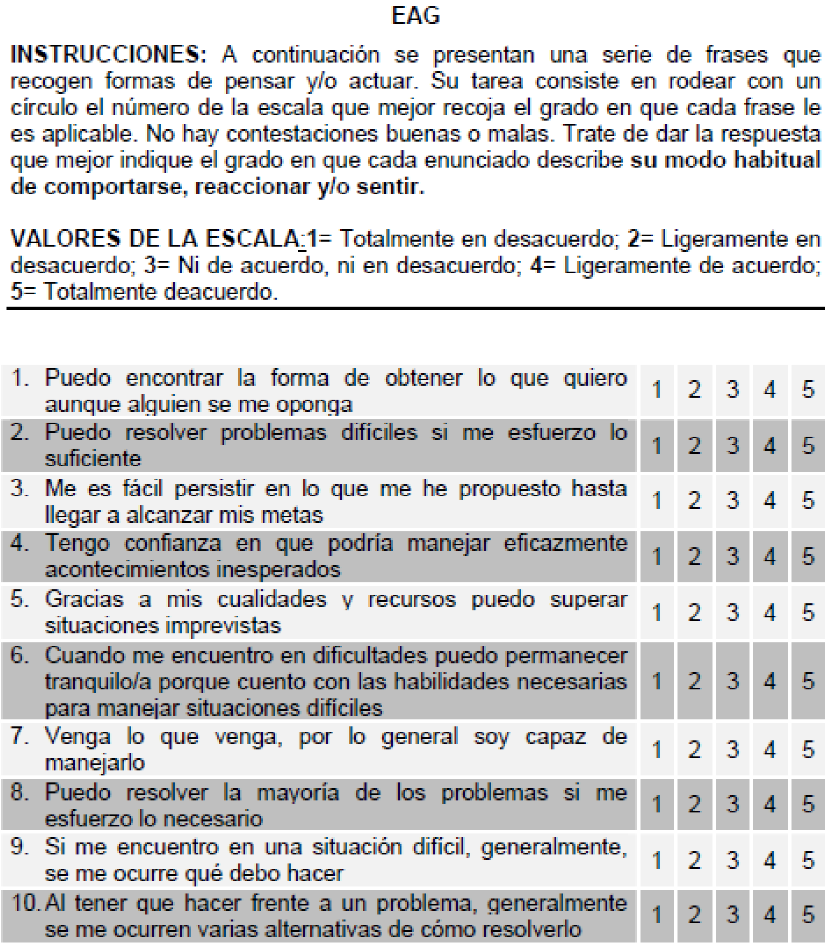
INSTRUCCIONES: Las siguientes frases pueden describirle a usted con mayor o menor precisión. Por ejemplo, el decir que usted es alguien "chistoso, a quien le gusta bromear", seguramente le describirá en mayor o menor medida. Por favor, para cada una de las siguientes frases, indique (redondeando el número correspondiente en la escala de la derecha) el grado en que está de acuerdo en que dicha frase le describe a usted.

VALORES DE LA ESCALA: 1= Muy en desacuerdo; 2= Ligeramente en desacuerdo; 3= Ni de acuerdo, ni en desacuerdo; 4= Ligeramente de acuerdo; 5= Muy de acuerdo.

Me veo a mí mismo/a como alguien que....

1. Es hablador	1	2	3	4	5
2. Tiende a criticar a los demás	1	2	3	4	5
3. Es minucioso en el trabajo	1	2	3	4	5
4. Es triste y melancólico	1	2	3	4	5
5. Es original, se le ocurren ideas nuevas	1	2	3	4	5
6. Es reservado	1	2	3	4	5
7. Es generoso y ayuda a los demás	1	2	3	4	5
8. A veces puede ser algo descuidado	1	2	3	4	5
9. Es tranquilo y controla bien el estrés	1	2	3	4	5
10. Tiene intereses muy diversos	1	2	3	4	5
11. Está lleno de energía	1	2	3	4	5
12. Prefiere trabajos rutinarios	1	2	3	4	5
13. Provoca disputas con los demás	1	2	3	4	5
14. Es un trabajador cumplidor, digno de confianza	1	2	3	4	5
15. Con frecuencia está tenso	1	2	3	4	5
16. Tiende a estar callado	1	2	3	4	5
17. Valora las experiencias artísticas y estéticas	1	2	3	4	5
18. Tiende a ser desorganizado	1	2	3	4	5
19. Es emocionalmente estable, difícil de alterar	1	2	3	4	5
20. Es imaginativo	1	2	3	4	5
21. Persevera hasta terminar el trabajo	1	2	3	4	5
22. A veces es maleducado, grosero, con los demás	1	2	3	4	5
23. Tiene inventiva	1	2	3	4	5
24. Generalmente se fía de los demás	1	2	3	4	5
25. Tiende a ser perezoso, vago	1	2	3	4	5
26. Se apura por cualquier cosa	1	2	3	4	5
27. A veces se muestra tímido y cohibido	1	2	3	4	5
28. Es indulgente, no le cuesta perdonar	1	2	3	4	5
29. Hace las cosas de manera eficiente	1	2	3	4	5
30. Tiene cambios de humor frecuentemente	1	2	3	4	5
31. Es ingenioso, intuitivo	1	2	3	4	5
32. Irradia, transmite, entusiasmo	1	2	3	4	5
33. A veces es frío y distante	1	2	3	4	5
34. Hace planes y los sigue escrupulosamente	1	2	3	4	5
35. Conserva la calma en las situaciones difíciles	1	2	3	4	5
36. Le gusta pensar, jugar con las ideas	1	2	3	4	5
37. Es considerado y amable con casi todo el mundo	1	2	3	4	5
38. Se pone nervioso fácilmente	1	2	3	4	5
39. Es entendido en arte, música o literatura	1	2	3	4	5
40. Es asertivo, no teme expresar claramente lo que desea	1	2	3	4	5
41. Le gusta cooperar con los demás	1	2	3	4	5
42. Se distrae con facilidad	1	2	3	4	5
43. Es extravertido, sociable	1	2	3	4	5
44. Tiene pocos intereses artísticos	1	2	3	4	5

Figure 73. Big Five Inventory Questionnaire.



EAG

INSTRUCCIONES: A continuación se presentan una serie de frases que recogen formas de pensar y/o actuar. Su tarea consiste en rodear con un círculo el número de la escala que mejor recoja el grado en que cada frase le es aplicable. No hay contestaciones buenas o malas. Trate de dar la respuesta que mejor indique el grado en que cada enunciado describe **su modo habitual de comportarse, reaccionar y/o sentir.**

VALORES DE LA ESCALA:1= Totalmente en desacuerdo; **2**= Ligeramente en desacuerdo; **3**= Ni de acuerdo, ni en desacuerdo; **4**= Ligeramente de acuerdo; **5**= Totalmente deacuerdo.

1. Puedo encontrar la forma de obtener lo que quiero aunque alguien se me oponga	1	2	3	4	5
2. Puedo resolver problemas difíciles si me esfuerzo lo suficiente	1	2	3	4	5
3. Me es fácil persistir en lo que me he propuesto hasta llegar a alcanzar mis metas	1	2	3	4	5
4. Tengo confianza en que podría manejar eficazmente acontecimientos inesperados	1	2	3	4	5
5. Gracias a mis cualidades y recursos puedo superar situaciones imprevistas	1	2	3	4	5
6. Cuando me encuentro en dificultades puedo permanecer tranquilo/a porque cuento con las habilidades necesarias para manejar situaciones difíciles	1	2	3	4	5
7. Venga lo que venga, por lo general soy capaz de manejarlo	1	2	3	4	5
8. Puedo resolver la mayoría de los problemas si me esfuerzo lo necesario	1	2	3	4	5
9. Si me encuentro en una situación difícil, generalmente, se me ocurre qué debo hacer	1	2	3	4	5
10. Al tener que hacer frente a un problema, generalmente se me ocurren varias alternativas de cómo resolverlo	1	2	3	4	5

Figure 74. General Self-Efficacy Scale Questionnaire.

13.2.4. Calibration questions

The following questions (similar to the ones used in polygraphs) were asked to the participants to calibrate the physiological data obtained.

- Question 1: Is Paris the capital of France?
- Question 2: Have you ever commited a mistake in your work or studies?
- Question 3: Is eight an even numbre?
- Question 4: Have you ever lied to your bosses or teachers in order to get some kind of benefit?
- Question 5: Is Gollum a fiction character?
- Question 6: Have you ever get advantage of other person’s work?
- Question 7: Have you ever taken something from a store without paying it?

13.2.5. IAPS pictures (calibration images)

The following 8 pictures from the IAPS data base were chosen.

- Picture 1: a dish on a table
- Picture 2: a book on a carpet.
- Picture 3: a group of nine people rafting.
- Picture 4: four people on a rollearcoaster screaming.
- Picture 5: a gun pointing at the viewer.
- Picture 6: a dog boofing.
- Picture 7: a little child with severe burns on his body.
- Picture 8: a hand with severe injuries with blood and a material going through the flesh.

13.2.6. IADS sounds (calibration sounds)

Alternatively to the IAPS images for calibration, the following 8 calibration sounds were used when participants were visually impared.

- Audio 1: Children’s Choir
- Audio 2: Birds singing
- Audio 3: Yawn (9seconds)
- Audio 4: Crowd celebrating (9 seconds)
- Audio 5: Rock music (9seconds)
- Audio 6: Female orgasm (9 seconds)
- Audio 7: Woman screaming (9 seconds)
- Audio 8: Woman screaming being beaten by a man (9 seconds)

13.2.7. Problems and logical series used in stage 1

The problems here presented were picked from the Activity 3: Ambient Intelligence: Affective automated tutor for the "everyday mathematics". Those used in the other two activities are similar to the ones reporte here.

13.2.7.a. *Task 1*

The 6 problems of task 1 are the following:

Problem 1: Arturo tiene tantos euros como indica el menor número de 3 cifras. Adela tiene tantos euros como indica el mayor número de 2 cifras. A uno de los dos amigos se le perdió un euro y entonces los dos se quedaron con la misma cantidad. ¿Quién perdió el euro?

- Arturo
- Adela
- Los dos
- Ninguno

Problem 2: Si Alicia se gastase 2 euros, le quedaría el doble de dinero que si se gastase 4 euros. ¿Cuántos euros tiene Alicia?

- 2
- 4
- 6
- 8

Problem 3: Antonio tiene en su corral 6 animales. Unos son vacas y otros son gallinas. Hoy le ha dado por averiguar las patas que tienen entre todos ellos y ha contado 16. ¿Cuántos animales son vacas y cuántos son gallinas?

- 4 vacas y 2 gallinas
- 2 vacas y 4 gallinas
- 2 vacas y 2 gallinas
- 4 vacas y 4 gallinas

Problem 4: Agustina tiene nueve monedas. Sólo una de ellas la tiene repetida. En total tiene 3 euros y 98 céntimos. ¿Cuál es la moneda que tiene repetida?

- 1 céntimo
- 2 céntimos
- 5 céntimos
- 10 céntimos
- 20 céntimos
- 50 céntimos
- 1 euro
- 2 euros

Problem 5: Un tren que ha salido de Badajoz hacia Madrid a las 10 de la mañana lleva una velocidad de 80 km/h. Media hora más tarde ha salido un tren de Madrid hacia Badajoz con una velocidad de 90 km/h. ¿A qué distancia estarán uno de otro justo una hora antes de cruzarse?

- 150 kilómetros
- 160 kilómetros
- 170 kilómetros
- 180 kilómetros

Problem 6: Argimiro es un gran aficionado a la pesca. Ayer pescó un pez de 9 kilos (no me preguntes de qué especie era, porque no lo sé). La cola pesaba la mitad que la cabeza y la cabeza pesaba 4 kilos menos que el cuerpo. ¿Cuántos kilos pesaba el cuerpo?

- 2
- 4
- 6
- 8

13.2.7.b. *Task 2*

The 6 problems of task 2 are the following:

Problem 1: En un juego de Trivial cada respuesta correcta puntúa con 5 puntos y cada respuesta incorrecta descuenta 2 puntos. ¿Qué puntuación se obtiene con 8 respuestas correctas y 4 incorrectas?

- 9
- 48
- 36
- 32

Problem 2: Los ingredientes de una receta de 40 magdalenas incluyen 400 gramos de mantequilla y 160 gramos de cerezas. ¿Qué cantidades de estos ingredientes se necesitarían para 10 magdalenas?

- 100 gramos de mantequilla y 80 gramos de cerezas
- 200 gramos de mantequilla y 80 gramos de cerezas
- 100 gramos de mantequilla y 40 gramos de cerezas
- 100 gramos de mantequilla y 60 gramos de cerezas

Problem 3: El siguiente número de la secuencia 4, 9, 19 es...

- 39
- 29
- 36

Problem 4: Samuel está cocinando una sopa de zanahorias siguiendo una receta. La receta está pensada para cuatro personas, pero quiere cocinar sopa suficiente para ocho.

A continuación presentamos los ingredientes para cuatro personas: 80 gramos de cebolla, 25 gramos de mantequilla, 1 diente de ajo, 400 gramos de zanahorias. ¿Qué ingredientes necesitará para ocho personas?

- 40 gramos de cebollas, 12,5 gramos de mantequilla, medio diente de ajo, 200 gramos de zanahorias
- 160 gramos de cebollas, 50 gramos de mantequilla, 2 dientes de ajo, 800 gramos de zanahorias
- 240 gramos de cebollas, 75 gramos de mantequilla, 3 dientes de ajo, 1,2 kilogramos de zanahorias
- 800 gramos de cebollas, 250 gramos de mantequilla, 10 dientes de ajo, 4000 gramos de zanahorias

Problem 5: Redondea cada uno de estos números a dos cifras decimales, y a continuación súmalos: 123,096, 54,882, 1,722, 15,907, 3,029. ¿Cuál es el total?

- 198,62
- 198,636
- 198,64
- 198,66

Problem 6: Si un hombre y medio beben una cerveza y media en un día y medio, ¿cuántas cervezas beberán seis hombres en seis días?

- 24
- 9
- 18
- 21

13.2.7.c. *Task 3*

The task 3 consisted in a of logical series. Typically, logical series involve a sequence of figures, which is not possible to do when participants are visually impaired. For that reason, equivalent logical series in numeric format were prepared for visually impared participants. In both cases, they were selected from available repositories in the literature.

13.2.7.c.i. Graphical logical series (people without seeing difficulties)

Problem 1:

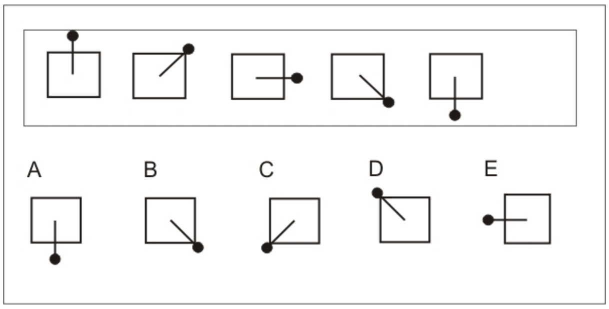


Figure 75. 1st Graphical logical series problem.

Problem 2:

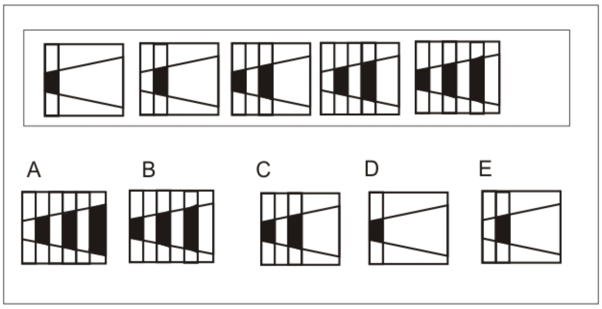


Figure 76. 2nd Graphical logical series problem.

Problem 3:

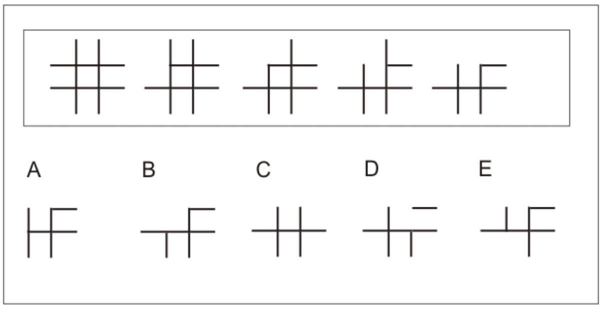


Figure 77. 3rd Graphical logical series problem.

Problem 4:

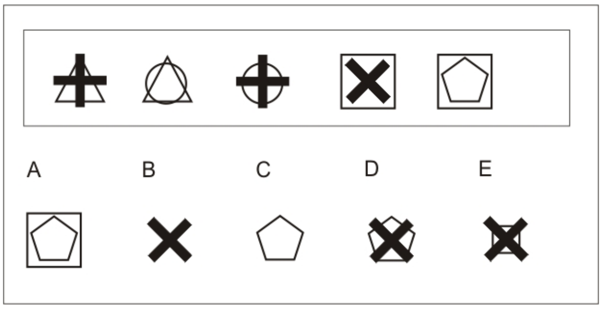


Figure 78. 4th Graphical logical series problem.

Problem 5:

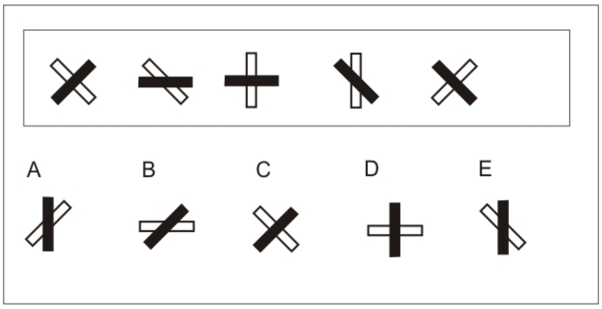


Figure 79. 5th Graphical logical series problem.

Problem 6:

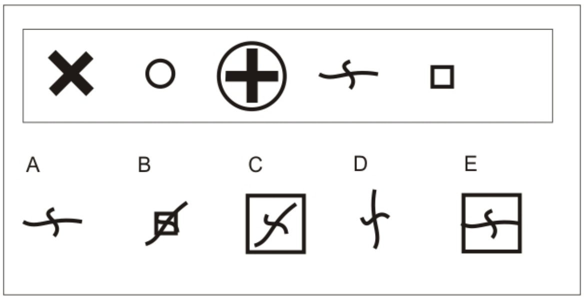


Figure 80. 6th Graphical logical series problem.

13.2.7.c.ii. Textual logical series (people with seeing difficulties)

Problem 1:

Las siguientes letras siguen una regla lógica ¿qué letra ha de completar la serie?

A C F J Ñ T A I ?

- M
- R

- Q
- E

Problem 2:

¿Qué número falta para completar la serie?

4 5 8 13 20 29 40 ?

- 48
- 42
- 53
- 36
- 56

Problem 3:

Las siguientes letras siguen una regla lógica ¿qué letra ha de completar la serie?

ñ ñ n ñ n m ñ n m l ñ n ?

- ñ
- m
- n
- l

Problem 4:

¿Qué número falta para completar la serie?

9 12 36 39 117 120 360 ?

- 180
- 352
- 245
- 363
- 179

Problem 5:

Las siguientes letras siguen una regla lógica ¿qué letra ha de completar la serie?

M O S V Z C G J ?

- K
- N
- L
- M

Problem 6:

¿Qué número o qué letra debemos poner en lugar de la interrogación para completar la serie?

? N 6 M 3 L

- 2
- 9
- 18
- 3
- E

13.2.8. Problems used in transition stage

The problems chosen for the transition stage were the following:

Problem 1: Ana y Miguel han ganado 36.000 de hacer los planos de un puente. Como no han trabajado el mismo tiempo se lo deben repartir de forma que a Ana le toquen cinco partes de lo que han ganado y a Miguel, siete partes. ¿Cuánto dinero le corresponde a Miguel?

Problem 2: Hemos mezclado 5 kilos de té de Tailandia, cuyo precio es de 4 euros el kilo, con 3 kilos de té de la India, que cuesta 6 euros el kilo. ¿Cuál será el precio de un kilo de mezcla?

Problem 3: Una motocicleta sale de una ciudad A hacia otra B a 40 km/h. Al mismo tiempo, un coche sale de B hacia A a una velocidad de 80 km/h. Si sabemos que la distancia entre A y B es de 300 km, ¿cuánto tiempo tardarán en encontrarse?

Problem 4: Pagué 1440,75 € por un ordenador después de obtener un descuento del 15% del precio marcado. ¿Cuál es el precio del ordenador sin descuento?

Problem 5: Un grifo de caudal constante vierte agua en un depósito cilíndrico. Se sabe que en 5 minutos el nivel del agua ha subido 20 cm. ¿Cuánto habría subido el nivel del agua en 13 minutos?

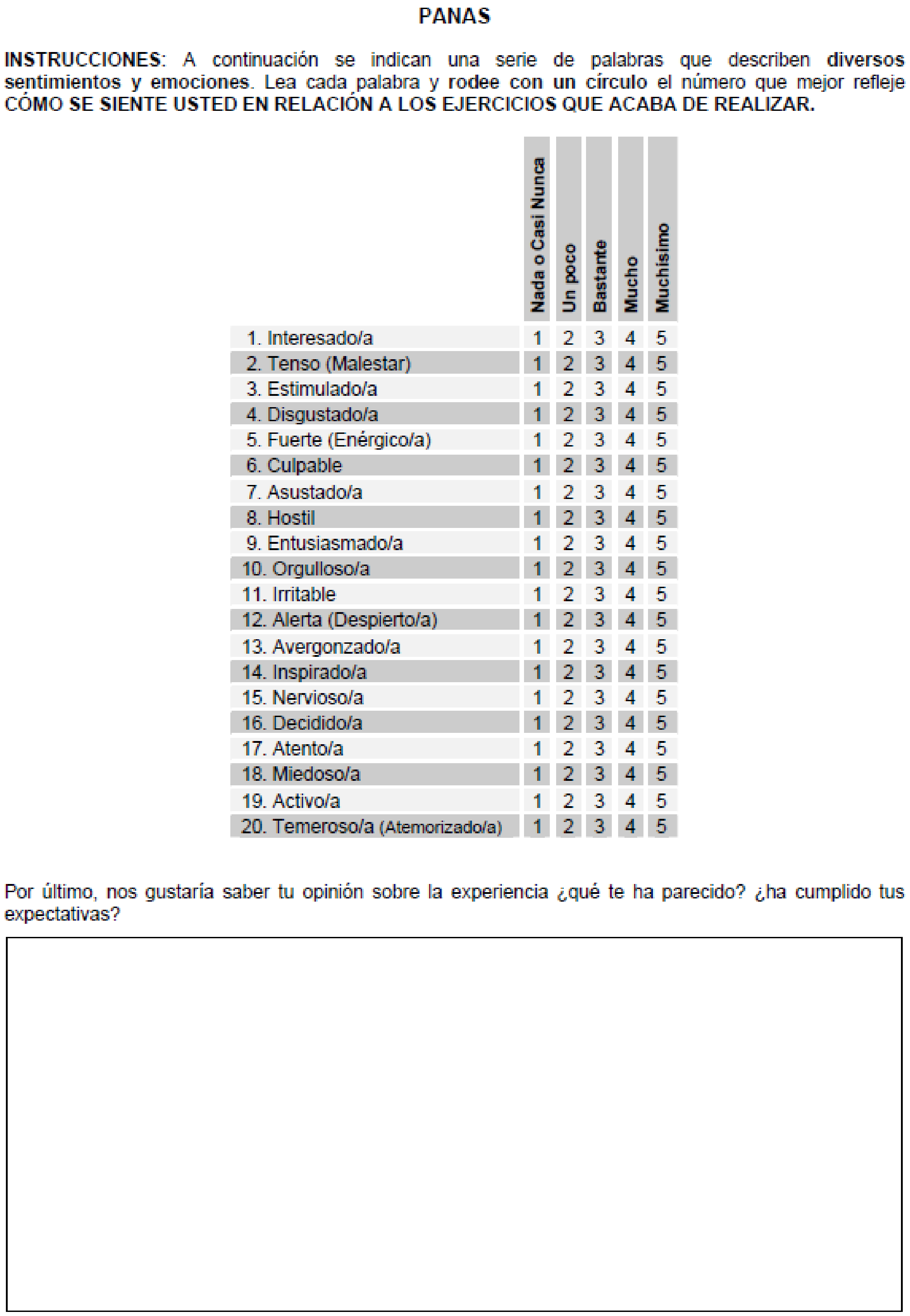
Problem 6: El agua que proviene de una acequia tarda 2 horas en llenar una balsa de 420 litros, mientras que la que entra por una tubería tarda 6.¿Cuánto tiempo tardará en llenarse la balsa si se abren la acequía y la tubería a la vez.

13.2.9. Proposed works for the essays in stage 2

The proposed words chosen for the essays to be written by participants in stage 2 were the following:

- 1st essay (participants can see the words during the task):
    - nature
    - global warming
    - greenhouse effect
    - ozone layer
    - climate change
- 2nd essay (participants have 30 seconds to memorize the words before starting the task):
    - TV
    - cartoon
    - drama
    - news
    - quiz show
- 3rd essay(participants have 30 seconds to memorize the words before starting the task):
    - dishonest
    - ringleader
    - rebellious
    - court
    - forger
    - prison

13.2.10. PANAS & Satisfaction questionnaire



PANAS

INSTRUCCIONES: A continuación se indican una serie de palabras que describen **diversos sentimientos y emociones**. Lea cada palabra y **rodee con un círculo** el número que mejor refleje CÓMO **SE SIENTE USTED EN RELACIÓN A LOS EJERCICIOS QUE ACABA DE REALIZAR.**

	Nada o Casi Nunca	Un poco	Bastante	Mucho	Muchísimo
1. Interesado/a	1	2	3	4	5
2. Tenso (Malestar)	1	2	3	4	5
3. Estimulado/a	1	2	3	4	5
4. Disgustado/a	1	2	3	4	5
5. Fuerte (Enérgico/a)	1	2	3	4	5
6. Culpable	1	2	3	4	5
7. Asustado/a	1	2	3	4	5
8. Hostil	1	2	3	4	5
9. Entusiasmado/a	1	2	3	4	5
10. Orgulloso/a	1	2	3	4	5
11. Irritable	1	2	3	4	5
12. Alerta (Despierto/a)	1	2	3	4	5
13. Avergonzado/a	1	2	3	4	5
14. Inspirado/a	1	2	3	4	5
15. Nervioso/a	1	2	3	4	5
16. Decidido/a	1	2	3	4	5
17. Atento/a	1	2	3	4	5
18. Miedoso/a	1	2	3	4	5
19. Activo/a	1	2	3	4	5
20. Temeroso/a (Atemorizado/a)	1	2	3	4	5

Por último, nos gustaría saber tu opinión sobre la experiencia ¿qué te ha parecido? ¿ha cumplido tus expectativas?

Figure 81. Positive and Negative Affect Schedule Questionnaire.

13.3. **Appendix III : Full results from the data mining processing in stage 1**

This section reports the full results obtained in the data mining process for the 735 models computed in stage 1. Two tables are reported for each category of models.

In the first table, for each of the 7 labelling approaches, the prediction results reporting accuracy and Coehen's Kappa for each of the prediction algorithm used and data source combination are ranked according to the score proposed in formula (4.6).

In the second table, the best prediction result per data source is reported, showing the ranking score, the accuracy, kappa, data source and algorithm used.

13.3.1. Approach 1: Valence given by the expert, with 10 years of experience in supporting learners in e-learning platforms.

Prediction results

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,70114943	-0,02260398	J48	Keyboard	0,68530066
0,68965517	-0,04446421	Bagging	Keyboard	0,6589902
0,62068966	-0,03758583	Random Forest	Keyboard	0,59736052
0,71264368	0	Bayesian Network	Keyboard	0,71264368
0,6091954	-0,17942584	Naïve Bayes	Keyboard	0,49989001
0,70114943	-0,02260398	SVM	Keyboard	0,68530066
0,71052632	-0,05025126	Neural Network	Keyboard	0,67482148
0,67816092	0,08283133	J48	Mouse	0,73433389
0,68965517	-0,01119242	Bagging	Mouse	0,68193626
0,68965517	0,21411843	Random Forest	Mouse	0,83732306
0,71264368	0	Bayesian Network	Mouse	0,71264368
0,49425287	-0,08013544	Naïve Bayes	Mouse	0,4546457
0,71264368	0	SVM	Mouse	0,71264368
0,67816092	-0,0656168	Neural Network	Mouse	0,63366217
0,67816092	-0,0656168	J48	Sentiment Analysis	0,63366217
0,71264368	0	Bagging	Sentiment Analysis	0,71264368
0,62068966	0,01509434	Random Forest	Sentiment Analysis	0,63005856
0,71264368	0	Bayesian Network	Sentiment Analysis	0,71264368
0,71264368	0	Naïve Bayes	Sentiment Analysis	0,71264368
0,71264368	0	SVM	Sentiment Analysis	0,71264368
0,63218391	-0,04819277	Neural Network	Sentiment Analysis	0,60171721
0,68965517	0,12773858	J48	Physiological	0,77775075
0,71264368	0,11978956	Bagging	Physiological	0,79801095
0,65517241	0,09312022	Random Forest	Physiological	0,71618221
0,71264368	0	Bayesian Network	Physiological	0,71264368
0,63218391	0,12287335	Naïve Bayes	Physiological	0,70986246
0,72413793	0,05605787	SVM	Physiological	0,76473156
0,70114943	-0,02260398	Neural Network	Physiological	0,68530066
0,70114943	0,04152542	J48	Keyboard + Mouse	0,73026495
0,66666667	-0,08609557	Bagging	Keyboard + Mouse	0,60926962
0,66666667	0,11193242	Random Forest	Keyboard + Mouse	0,74128828
0,71264368	0	Bayesian Network	Keyboard + Mouse	0,71264368
0,54022989	-0,24108417	Naïve Bayes	Keyboard + Mouse	0,40998901
0,70114943	-0,02260398	SVM	Keyboard + Mouse	0,68530066

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,72368421	-0,02570694	Neural Network	Keyboard + Mouse	0,7050805
0,67816092	-0,0656168	J48	Keyboard + Sentiment Analysis	0,63366217
0,68965517	-0,01119242	Bagging	Keyboard + Sentiment Analysis	0,68193626
0,70114943	0,12393493	Random Forest	Keyboard + Sentiment Analysis	0,78804633
0,71264368	0	Bayesian Network	Keyboard + Sentiment Analysis	0,71264368
0,63218391	-0,14379622	Naïve Bayes	Keyboard + Sentiment Analysis	0,54127825
0,70114943	-0,02260398	SVM	Keyboard + Sentiment Analysis	0,68530066
0,73684211	0	Neural Network	Keyboard + Sentiment Analysis	0,73684211
0,65517241	0,01731928	J48	Keyboard + Physiological	0,66651953
0,71264368	0,06371072	Bagging	Keyboard + Physiological	0,75804672
0,6091954	-0,11370482	Random Forest	Keyboard + Physiological	0,53992695
0,71264368	0	Bayesian Network	Keyboard + Physiological	0,71264368
0,65517241	-0,04066986	Naïve Bayes	Keyboard + Physiological	0,62852665
0,72413793	0,11525424	SVM	Keyboard + Physiological	0,8075979
0,72368421	-0,02570694	Neural Network	Keyboard + Physiological	0,7050805
0,66666667	0,03665521	J48	Mouse + Sentiment Analysis	0,69110347
0,71264368	0,03290351	Bagging	Mouse + Sentiment Analysis	0,73609216
0,67816092	0,10769231	Random Forest	Mouse + Sentiment Analysis	0,75119363
0,71264368	0	Bayesian Network	Mouse + Sentiment Analysis	0,71264368
0,49425287	-0,12720848	Naïve Bayes	Mouse + Sentiment Analysis	0,43137972
0,71264368	0	SVM	Mouse + Sentiment Analysis	0,71264368
0,67816092	-0,0656168	Neural Network	Mouse + Sentiment Analysis	0,63366217
0,62068966	0,1260274	J48	Mouse + Physiological	0,69891356
0,70114943	0,01049869	Bagging	Mouse + Physiological	0,70851057
0,63218391	-0,01978022	Random Forest	Mouse + Physiological	0,61967917
0,71264368	0	Bayesian Network	Mouse + Physiological	0,71264368
0,5862069	0,15167931	Naïve Bayes	Mouse + Physiological	0,67512235
0,72413793	0,08661417	SVM	Mouse + Physiological	0,78685854
0,66666667	-0,08609557	Neural Network	Mouse + Physiological	0,60926962

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,68965517	0,12773858	J48	Sentiment Analysis + Physiological	0,77775075
0,71264368	0,06371072	Bagging	Sentiment Analysis + Physiological	0,75804672
0,67816092	0,13124108	Random Forest	Sentiment Analysis + Physiological	0,76716349
0,71264368	0	Bayesian Network	Sentiment Analysis + Physiological	0,71264368
0,59770115	0,05169729	Naïve Bayes	Sentiment Analysis + Physiological	0,62860068
0,72413793	0,05605787	SVM	Sentiment Analysis + Physiological	0,76473156
0,71264368	0	Neural Network	Sentiment Analysis + Physiological	0,71264368
0,64367816	-0,05972495	J48	Keyboard + Mouse + Sentiment Analysis	0,60523451
0,71264368	0,06371072	Bagging	Keyboard + Mouse + Sentiment Analysis	0,75804672
0,59770115	-0,10046982	Random Forest	Keyboard + Mouse + Sentiment Analysis	0,53765022
0,71264368	0	Bayesian Network	Keyboard + Mouse + Sentiment Analysis	0,71264368
0,56321839	-0,17902996	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis	0,46238543
0,70114943	-0,02260398	SVM	Keyboard + Mouse + Sentiment Analysis	0,68530066
0,72368421	-0,02570694	Neural Network	Keyboard + Mouse + Sentiment Analysis	0,7050805
0,64367816	0,1194907	J48	Keyboard + Mouse + Physiological	0,72059171
0,66666667	-0,05256571	Bagging	Keyboard + Mouse + Physiological	0,63162286
0,66666667	0,03665521	Random Forest	Keyboard + Mouse + Physiological	0,69110347
0,71264368	0	Bayesian Network	Keyboard + Mouse + Physiological	0,71264368
0,63218391	0,05691057	Naïve Bayes	Keyboard + Mouse + Physiological	0,66816185
0,67816092	-0,03220339	SVM	Keyboard + Mouse + Physiological	0,65632184
0,72368421	-0,02570694	Neural Network	Keyboard + Mouse + Physiological	0,7050805
0,56321839	-0,09253139	J48	Keyboard + Sentiment Analysis + Physiological	0,51110301
0,67816092	-8,22E+11	Bagging	Keyboard + Sentiment Analysis + Physiological	-5,5724E+11
0,64367816	0,05068638	Random Forest	Keyboard + Sentiment Analysis + Physiological	0,67630388

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,68965517	-0,04446421	Bayesian Network	Keyboard + Sentiment Analysis + Physiological	0,6589902
0,65517241	-0,1059322	Naïve Bayes	Keyboard + Sentiment Analysis + Physiological	0,58576856
0,71264368	0,06371072	SVM	Keyboard + Sentiment Analysis + Physiological	0,75804672
0,72368421	-0,02570694	Neural Network	Keyboard + Sentiment Analysis + Physiological	0,7050805
0,66666667	0,13447684	J48	Mouse + Sentiment Analysis + Physiological	0,7563179
0,68965517	0,07701375	Bagging	Mouse + Sentiment Analysis + Physiological	0,74276811
0,66666667	0,06312662	Random Forest	Mouse + Sentiment Analysis + Physiological	0,70875108
0,71264368	0	Bayesian Network	Mouse + Sentiment Analysis + Physiological	0,71264368
0,54022989	-0,00288184	Naïve Bayes	Mouse + Sentiment Analysis + Physiological	0,53867303
0,72413793	0,11525424	SVM	Mouse + Sentiment Analysis + Physiological	0,8075979
0,72413793	0,05605787	Neural Network	Mouse + Sentiment Analysis + Physiological	0,76473156
0,65517241	0,06918688	J48	Keyboard + Mouse + Sentiment Analysis + Physiological	0,70050175
0,74712644	0,18898305	Bagging	Keyboard + Mouse + Sentiment Analysis + Physiological	0,88832067
0,71264368	0,1453831	Random Forest	Keyboard + Mouse + Sentiment Analysis + Physiological	0,81625003
0,71264368	0	Bayesian Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,71264368
0,62068966	-0,01055966	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis + Physiological	0,61413538
0,72413793	0,14215283	SVM	Keyboard + Mouse + Sentiment Analysis + Physiological	0,82707619

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,69354839	-0,03152364	Neural Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,67168522

Table 29. Prediction results for labeling approach 1

Best prediction result per data source

Score	Accuracy	Kappa	Data Source	Algorithm
1,34363316	0,84761905	0,58518519	Sentiment Analysis	Random Forest
1,30204082	0,82857143	0,57142857	Mouse + Sentiment Analysis + Physiological	J48
1,27047619	0,82857143	0,53333333	Keyboard + Mouse + Sentiment Analysis	Bagging
1,18736842	0,8	0,48421053	Keyboard + Sentiment Analysis	J48
1,13988958	0,79047619	0,44202899	Mouse + Sentiment Analysis	J48
1,13049586	0,76190476	0,48377581	Sentiment Analysis + Physiological	Bayesian Network
1,09231006	0,76190476	0,43365696	Keyboard + Mouse + Sentiment Analysis + Physiological	Bayesian Network
1,08981241	0,75238095	0,44848485	Keyboard + Sentiment Analysis + Physiological	Bayesian Network
0,78754579	0,71428571	0,1025641	Keyboard + Physiological	Bagging
0,77248677	0,6952381	0,11111111	Physiological	Random Forest
0,74747475	0,7047619	0,06060606	Mouse	Bagging
0,72150638	0,63809524	0,13071895	Mouse + Physiological	Naïve Bayes
0,71428571	0,71428571	0	Keyboard	Bayesian Network
0,71428571	0,71428571	0	Keyboard	Bayesian Network
0,69818041	0,67619048	0,03252033	Keyboard + Mouse + Physiological	Random Forest

Table 30. Best prediction per data source for labeling approach 1

13.3.2. Approach 2: Valence given by two psychologist, with experience in motivational and emotional issues.

Prediction results

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,73170732	0,30077519	J48	Keyboard	0,95178673
0,80487805	0,44781145	Bagging	Keyboard	1,16531165
0,68292683	0,21502209	Random Forest	Keyboard	0,82977118
0,68292683	0,01841621	Bayesian Network	Keyboard	0,69550375
0,6097561	0,15463918	Naïve Bayes	Keyboard	0,70404828
0,70731707	0	SVM	Keyboard	0,70731707
0,43243243	0,13666667	Neural Network	Keyboard	0,49153153
0,70731707	0,21656051	J48	Mouse	0,86049402
0,73170732	0,1694291	Bagging	Mouse	0,85567983
0,68292683	0,21502209	Random Forest	Mouse	0,82977118
0,65853659	-0,09125475	Bayesian Network	Mouse	0,59844199
0,75609756	0,32343234	Naïve Bayes	Mouse	1,00064397
0,80487805	0,41428571	SVM	Mouse	1,13832753
0,2195122	-0,02260327	Neural Network	Mouse	0,2145505
0,68292683	0,12765957	J48	Sentiment Analysis	0,77010898
0,63414634	-0,00654664	Bagging	Sentiment Analysis	0,62999481
0,70731707	0,21656051	Random Forest	Sentiment Analysis	0,86049402
0,65853659	-0,09125475	Bayesian Network	Sentiment Analysis	0,59844199
0,65853659	0,08598726	Naïve Bayes	Sentiment Analysis	0,71516234
0,70731707	0	SVM	Sentiment Analysis	0,70731707
0,70731707	0,21656051	Neural Network	Sentiment Analysis	0,86049402
0,56097561	-0,24242424	J48	Physiological	0,42498152
0,68292683	0,0762565	Bagging	Physiological	0,73500444
0,75609756	0,30976431	Random Forest	Physiological	0,9903096
0,70731707	0	Bayesian Network	Physiological	0,70731707
0,53658537	0,11778029	Naïve Bayes	Physiological	0,59978455
0,70731707	0	SVM	Physiological	0,70731707
0,70731707	0	Neural Network	Physiological	0,70731707
0,65853659	0,13293051	J48	Keyboard + Mouse	0,74607619

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,75609756	0,30976431	Bagging	Keyboard + Mouse	0,9903096
0,70731707	0,25679758	Random Forest	Keyboard + Mouse	0,88895439
0,65853659	-0,09125475	Bayesian Network	Keyboard + Mouse	0,59844199
0,73170732	0,27608347	Naïve Bayes	Keyboard + Mouse	0,93371961
0,80487805	0,41428571	SVM	Keyboard + Mouse	1,13832753
0,51351351	0,19759036	Neural Network	Keyboard + Mouse	0,61497883
0,68292683	0,17364341	J48	Keyboard + Sentiment Analysis	0,80151257
0,80487805	0,47770701	Bagging	Keyboard + Sentiment Analysis	1,18937393
0,85365854	0,60828025	Random Forest	Keyboard + Sentiment Analysis	1,37292217
0,63414634	-0,13259669	Bayesian Network	Keyboard + Sentiment Analysis	0,55006064
0,73170732	0,39625167	Naïve Bayes	Keyboard + Sentiment Analysis	1,02164757
0,70731707	0,06463878	SVM	Keyboard + Sentiment Analysis	0,75303719
0,35135135	-0,03738318	Neural Network	Keyboard + Sentiment Analysis	0,33821672
0,65853659	0,13293051	J48	Keyboard + Physiological	0,74607619
0,7804878	0,36048527	Bagging	Keyboard + Physiological	1,06184216
0,75609756	0,30976431	Random Forest	Keyboard + Physiological	0,9903096
0,65853659	-0,09125475	Bayesian Network	Keyboard + Physiological	0,59844199
0,65853659	0,28070175	Naïve Bayes	Keyboard + Physiological	0,84338896
0,68292683	-0,04715128	SVM	Keyboard + Physiological	0,65072596
0,27027027	-0,0111336	Neural Network	Keyboard + Physiological	0,26726119

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,70731707	0,21656051	J48	Mouse + Sentiment Analysis	0,86049402
0,75609756	0,30976431	Bagging	Mouse + Sentiment Analysis	0,9903096
0,80487805	0,44781145	Random Forest	Mouse + Sentiment Analysis	1,16531165
0,70731707	0	Bayesian Network	Mouse + Sentiment Analysis	0,70731707
0,7804878	0,36048527	Naïve Bayes	Mouse + Sentiment Analysis	1,06184216
0,80487805	0,41428571	SVM	Mouse + Sentiment Analysis	1,13832753
0,48780488	0	Neural Network	Mouse + Sentiment Analysis	0,48780488
0,73170732	0,26186579	J48	Mouse + Physiological	0,92331643
0,73170732	0,26186579	Bagging	Mouse + Physiological	0,92331643
0,73170732	0,26186579	Random Forest	Mouse + Physiological	0,92331643
0,70731707	0	Bayesian Network	Mouse + Physiological	0,70731707
0,68292683	0,17364341	Naïve Bayes	Mouse + Physiological	0,80151257
0,7804878	0,36048527	SVM	Mouse + Physiological	1,06184216
0,26829268	0,07099698	Neural Network	Mouse + Physiological	0,28734065
0,80487805	0,52873563	J48	Sentiment Analysis + Physiological	1,23044575
0,82926829	0,50259965	Bagging	Sentiment Analysis + Physiological	1,24605825
0,75609756	0,30976431	Random Forest	Sentiment Analysis + Physiological	0,9903096
0,70731707	0	Bayesian Network	Sentiment Analysis + Physiological	0,70731707

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,58536585	0,17903416	Naïve Bayes	Sentiment Analysis + Physiological	0,69016634
0,68292683	-0,04715128	SVM	Sentiment Analysis + Physiological	0,65072596
0,70731707	0	Neural Network	Sentiment Analysis + Physiological	0,70731707
0,65853659	0,13293051	J48	Keyboard + Mouse + Sentiment Analysis	0,74607619
0,73170732	0,21837088	Bagging	Keyboard + Mouse + Sentiment Analysis	0,89149089
0,82926829	0,50259965	Random Forest	Keyboard + Mouse + Sentiment Analysis	1,24605825
0,68292683	0,01841621	Bayesian Network	Keyboard + Mouse + Sentiment Analysis	0,69550375
0,7804878	0,39607201	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis	1,08961718
0,80487805	0,41428571	SVM	Keyboard + Mouse + Sentiment Analysis	1,13832753
0,51351351	0,06591865	Neural Network	Keyboard + Mouse + Sentiment Analysis	0,54736363
0,58536585	0,02244039	J48	Keyboard + Mouse + Physiological	0,59850169
0,7804878	0,36048527	Bagging	Keyboard + Mouse + Physiological	1,06184216
0,7804878	0,39607201	Random Forest	Keyboard + Mouse + Physiological	1,08961718

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,70731707	0,06463878	Bayesian Network	Keyboard + Mouse + Physiological	0,75303719
0,65853659	0,21369863	Naïve Bayes	Keyboard + Mouse + Physiological	0,79926495
0,7804878	0,36048527	SVM	Keyboard + Mouse + Physiological	1,06184216
0,32432432	0,0522541	Neural Network	Keyboard + Mouse + Physiological	0,3412716
0,65853659	0,13293051	J48	Keyboard + Sentiment Analysis + Physiological	0,74607619
0,7804878	0,42790698	Bagging	Keyboard + Sentiment Analysis + Physiological	1,11446398
0,68292683	0,21502209	Random Forest	Keyboard + Sentiment Analysis + Physiological	0,82977118
0,58536585	-0,20797227	Bayesian Network	Keyboard + Sentiment Analysis + Physiological	0,46362599
0,65853659	0,28070175	Naïve Bayes	Keyboard + Sentiment Analysis + Physiological	0,84338896
0,7804878	0,42790698	SVM	Keyboard + Sentiment Analysis + Physiological	1,11446398
0,27027027	0,01576355	Neural Network	Keyboard + Sentiment Analysis + Physiological	0,27453069
0,73170732	0,30077519	J48	Mouse + Sentiment Analysis + Physiological	0,95178673

Accuracy	**Coehn's Kappa**	**Predictor**	**Data Sources**	**Score**
0,75609756	0,30976431	Bagging	Mouse + Sentiment Analysis + Physiological	0,9903096
0,7804878	0,39607201	Random Forest	Mouse + Sentiment Analysis + Physiological	1,08961718
0,70731707	0	Bayesian Network	Mouse + Sentiment Analysis + Physiological	0,70731707
0,7804878	0,42790698	Naïve Bayes	Mouse + Sentiment Analysis + Physiological	1,11446398
0,80487805	0,41428571	SVM	Mouse + Sentiment Analysis + Physiological	1,13832753
0,41463415	0,05110897	Neural Network	Mouse + Sentiment Analysis + Physiological	0,43582567
0,80487805	0,47770701	J48	Keyboard + Mouse + Sentiment Analysis + Physiological	1,18937393
0,75609756	0,30976431	Bagging	Keyboard + Mouse + Sentiment Analysis + Physiological	0,9903096
0,68292683	0,17364341	Random Forest	Keyboard + Mouse + Sentiment Analysis + Physiological	0,80151257
0,73170732	0,30077519	Bayesian Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,95178673

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,82926829	0,53027823	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis + Physiological	1,26901122
0,80487805	0,47770701	SVM	Keyboard + Mouse + Sentiment Analysis + Physiological	1,18937393
0,5	0,18706048	Neural Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,59353024

Table 31. Prediction results for labeling approach 2

Best prediction result per data source

Score	Accuracy	Kappa	Data Source	Algorithm
1,372922169	0,85365854	0,60828025	Keyboard + Sentiment Analysis	Random Forest
1,269011217	0,82926829	0,53027823	Keyboard + Mouse + Sentiment Analysis + Physiological	Naïve Bayes
1,246058249	0,82926829	0,50259965	Sentiment Analysis + Physiological	Bagging
1,246058249	0,82926829	0,50259965	Sentiment Analysis + Physiological	Bagging
1,165311653	0,80487805	0,44781145	Keyboard	Bagging
1,165311653	0,80487805	0,44781145	Keyboard	Bagging
1,138327526	0,80487805	0,41428571	Mouse	SVM
1,138327526	0,80487805	0,41428571	Mouse	SVM
1,138327526	0,80487805	0,41428571	Mouse	SVM
1,114463982	0,7804878	0,42790698	Keyboard + Sentiment Analysis + Physiological	Bagging
1,089617181	0,7804878	0,39607201	Keyboard + Mouse + Physiological	Random Forest
1,061842161	0,7804878	0,36048527	Keyboard + Physiological	Bagging
1,061842161	0,7804878	0,36048527	Keyboard + Physiological	Bagging
0,9903096	0,75609756	0,30976431	Physiological	Random Forest
0,860494019	0,70731707	0,21656051	Sentiment Analysis	Random Forest

Table 32. Best prediction per data source for labeling approach 2

13.3.3. Approcah 3: Arousal given by two psychologist, with experience in motivational and emotional issues.

Prediction results

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,9122807	0	J48	Keyboard	0,9122807
0,9122807	0	Bagging	Keyboard	0,9122807
0,89473684	-0,03012048	Random Forest	Keyboard	0,86778694
0,9122807	0	Bayesian Network	Keyboard	0,9122807
0,8245614	-0,08571429	Naïve Bayes	Keyboard	0,75388471
0,9122807	0	SVM	Keyboard	0,9122807
0,16326531	-0,01209068	Neural Network	Keyboard	0,16129132
0,89473684	-0,03012048	J48	Mouse	0,86778694
0,9122807	0	Bagging	Mouse	0,9122807
0,9122807	0	Random Forest	Mouse	0,9122807
0,9122807	0	Bayesian Network	Mouse	0,9122807
0,59649123	-0,15915119	Naïve Bayes	Mouse	0,50155894
0,9122807	0	SVM	Mouse	0,9122807
0,22807018	-0,0754717	Neural Network	Mouse	0,21085733
0,9122807	0	J48	Sentiment Analysis	0,9122807
0,9122807	0	Bagging	Sentiment Analysis	0,9122807
0,9122807	0	Random Forest	Sentiment Analysis	0,9122807
0,9122807	0	Bayesian Network	Sentiment Analysis	0,9122807
0,9122807	0	Naïve Bayes	Sentiment Analysis	0,9122807
0,9122807	0	SVM	Sentiment Analysis	0,9122807
0,87719298	-0,05277045	Neural Network	Sentiment Analysis	0,83090312
0,9122807	0,24802111	J48	Physiological	1,13854557
0,9122807	0	Bagging	Physiological	0,9122807
0,9122807	0,24802111	Random Forest	Physiological	1,13854557
0,89473684	-0,03012048	Bayesian Network	Physiological	0,86778694
0,8245614	-0,09615385	Naïve Bayes	Physiological	0,74527665
0,9122807	0	SVM	Physiological	0,9122807
0,9122807	0	Neural Network	Physiological	0,9122807
0,9122807	0	J48	Keyboard + Mouse	0,9122807

Accuracy	**Coehn's Kappa**	**Predictor**	**Data Sources**	**Score**
0,9122807	0	Bagging	Keyboard + Mouse	0,9122807
0,9122807	0	Random Forest	Keyboard + Mouse	0,9122807
0,9122807	0	Bayesian Network	Keyboard + Mouse	0,9122807
0,52631579	-0,16238671	Naïve Bayes	Keyboard + Mouse	0,4408491
0,9122807	0	SVM	Keyboard + Mouse	0,9122807
0,55102041	0,01100917	Neural Network	Keyboard + Mouse	0,55708669
0,89473684	-0,03012048	J48	Keyboard + Sentiment Analysis	0,86778694
0,9122807	0	Bagging	Keyboard + Sentiment Analysis	0,9122807
0,9122807	0	Random Forest	Keyboard + Sentiment Analysis	0,9122807
0,9122807	0	Bayesian Network	Keyboard + Sentiment Analysis	0,9122807
0,8245614	-0,09615385	Naïve Bayes	Keyboard + Sentiment Analysis	0,74527665
0,9122807	0	SVM	Keyboard + Sentiment Analysis	0,9122807
0,34693878	0,04390244	Neural Network	Keyboard + Sentiment Analysis	0,36217023
0,89473684	0,1971831	J48	Keyboard + Physiological	1,07116383
0,9122807	0	Bagging	Keyboard + Physiological	0,9122807
0,89473684	-0,03012048	Random Forest	Keyboard + Physiological	0,86778694
0,9122807	0	Bayesian Network	Keyboard + Physiological	0,9122807
0,84210526	-0,0845666	Naïve Bayes	Keyboard + Physiological	0,77089129
0,9122807	0	SVM	Keyboard + Physiological	0,9122807
0,12244898	0,01079812	Neural Network	Keyboard + Physiological	0,1237712

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,9122807	0	J48	Mouse + Sentiment Analysis	0,9122807
0,9122807	0	Bagging	Mouse + Sentiment Analysis	0,9122807
0,87719298	-0,05277045	Random Forest	Mouse + Sentiment Analysis	0,83090312
0,9122807	0	Bayesian Network	Mouse + Sentiment Analysis	0,9122807
0,70175439	-0,14134276	Naïve Bayes	Mouse + Sentiment Analysis	0,60256649
0,9122807	0	SVM	Mouse + Sentiment Analysis	0,9122807
0,21052632	-0,07818411	Neural Network	Mouse + Sentiment Analysis	0,1940665
0,8245614	-0,09615385	J48	Mouse + Physiological	0,74527665
0,9122807	0	Bagging	Mouse + Physiological	0,9122807
0,9122807	0	Random Forest	Mouse + Physiological	0,9122807
0,89473684	-0,03012048	Bayesian Network	Mouse + Physiological	0,86778694
0,75438596	-0,12711864	Naïve Bayes	Mouse + Physiological	0,65848944
0,9122807	0	SVM	Mouse + Physiological	0,9122807
0,40350877	0,04059406	Neural Network	Mouse + Physiological	0,41988883
0,87719298	0,1564482	J48	Sentiment Analysis + Physiological	1,01442825
0,9122807	0	Bagging	Sentiment Analysis + Physiological	0,9122807
0,89473684	-0,03012048	Random Forest	Sentiment Analysis + Physiological	0,86778694
0,9122807	0	Bayesian Network	Sentiment Analysis + Physiological	0,9122807

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,87719298	0,1564482	Naïve Bayes	Sentiment Analysis + Physiological	1,01442825
0,9122807	0	SVM	Sentiment Analysis + Physiological	0,9122807
0,9122807	0	Neural Network	Sentiment Analysis + Physiological	0,9122807
0,87719298	-0,05277045	J48	Keyboard + Mouse + Sentiment Analysis	0,83090312
0,9122807	0	Bagging	Keyboard + Mouse + Sentiment Analysis	0,9122807
0,9122807	0	Random Forest	Keyboard + Mouse + Sentiment Analysis	0,9122807
0,9122807	0	Bayesian Network	Keyboard + Mouse + Sentiment Analysis	0,9122807
0,70175439	-0,14134276	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis	0,60256649
0,9122807	0	SVM	Keyboard + Mouse + Sentiment Analysis	0,9122807
0,3877551	-0,04850214	Neural Network	Keyboard + Mouse + Sentiment Analysis	0,36894815
0,87719298	-0,05277045	J48	Keyboard + Mouse + Physiological	0,83090312
0,9122807	0	Bagging	Keyboard + Mouse + Physiological	0,9122807
0,9122807	0	Random Forest	Keyboard + Mouse + Physiological	0,9122807

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,89473684	-0,03012048	Bayesian Network	Keyboard + Mouse + Physiological	0,86778694
0,70175439	-0,14134276	Naïve Bayes	Keyboard + Mouse + Physiological	0,60256649
0,9122807	0	SVM	Keyboard + Mouse + Physiological	0,9122807
0,46938776	0,13037543	Neural Network	Keyboard + Mouse + Physiological	0,53058438
0,87719298	-0,05277045	J48	Keyboard + Sentiment Analysis + Physiological	0,83090312
0,9122807	0	Bagging	Keyboard + Sentiment Analysis + Physiological	0,9122807
0,9122807	0	Random Forest	Keyboard + Sentiment Analysis + Physiological	0,9122807
0,9122807	0	Bayesian Network	Keyboard + Sentiment Analysis + Physiological	0,9122807
0,8245614	-0,09615385	Naïve Bayes	Keyboard + Sentiment Analysis + Physiological	0,74527665
0,9122807	0	SVM	Keyboard + Sentiment Analysis + Physiological	0,9122807
0,28571429	0,01152738	Neural Network	Keyboard + Sentiment Analysis + Physiological	0,28900782
0,85964912	0,12307692	J48	Mouse + Sentiment Analysis + Physiological	0,96545209

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,9122807	0	Bagging	Mouse + Sentiment Analysis + Physiological	0,9122807
0,9122807	0	Random Forest	Mouse + Sentiment Analysis + Physiological	0,9122807
0,9122807	0	Bayesian Network	Mouse + Sentiment Analysis + Physiological	0,9122807
0,77192982	-0,12102874	Naïve Bayes	Mouse + Sentiment Analysis + Physiological	0,67850413
0,9122807	0	SVM	Mouse + Sentiment Analysis + Physiological	0,9122807
0,31578947	-0,04562559	Neural Network	Mouse + Sentiment Analysis + Physiological	0,30138139
0,85964912	-0,07042254	J48	Keyboard + Mouse + Sentiment Analysis + Physiological	0,79911045
0,9122807	0	Bagging	Keyboard + Mouse + Sentiment Analysis + Physiological	0,9122807
0,9122807	0	Random Forest	Keyboard + Mouse + Sentiment Analysis + Physiological	0,9122807
0,89473684	-0,03012048	Bayesian Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,86778694

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,78947368	-0,11400651	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis + Physiological	0,69946854
0,87719298	-0,05277045	SVM	Keyboard + Mouse + Sentiment Analysis + Physiological	0,83090312
0,31818182	-0,06365834	Neural Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,29792689

Table 33. Prediction results for labeling approach 3

Best prediction result per data source

Score	Accuracy	Kappa	Data Source	Algorithm
1,13854557	0,9122807	0,24802111	Physiological	J48
1,07116383	0,89473684	0,1971831	Keyboard + Physiological	J48
1,01442825	0,87719298	0,1564482	Sentiment Analysis + Physiological	J48
0,96545209	0,85964912	0,12307692	Mouse + Sentiment Analysis + Physiological	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48
0,9122807	0,9122807	0	Keyboard	J48

Table 34. Best prediction per data source for labeling approach 3

13.3.4. Approach 4: Mean SAM valence values given by participants during the problems in each task.

Prediction results

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,53846154	-0,0609358	J48	Keyboard	0,50564995
0,46153846	-0,1530664	Bagging	Keyboard	0,39089243
0,43076923	-0,18648249	Random Forest	Keyboard	0,35043831
0,56923077	0	Bayesian Network	Keyboard	0,56923077
0,35384615	-0,29506641	Naïve Bayes	Keyboard	0,24943804
0,53846154	-0,0609358	SVM	Keyboard	0,50564995
0,30909091	-0,08571429	Neural Network	Keyboard	0,2825974
0,55384615	-0,01072386	J48	Mouse	0,54790678
0,44615385	-0,18062563	Bagging	Mouse	0,36556703
0,49230769	-0,03075444	Random Forest	Mouse	0,47716704
0,56923077	0	Bayesian Network	Mouse	0,56923077
0,47692308	0,00270758	Naïve Bayes	Mouse	0,47821438
0,44615385	-0,21369295	SVM	Mouse	0,35081392
0,41538462	0,15091097	Neural Network	Mouse	0,47807071
0,6	0,19138756	J48	Sentiment Analysis	0,71483254
0,64615385	0,26245683	Bagging	Sentiment Analysis	0,81574134
0,66153846	0,285	Random Forest	Sentiment Analysis	0,85007692
0,56923077	0	Bayesian Network	Sentiment Analysis	0,56923077
0,61538462	0,23887588	Naïve Bayes	Sentiment Analysis	0,76238516
0,58461538	0,04045927	SVM	Sentiment Analysis	0,60826849
0,55384615	0,07005427	Neural Network	Sentiment Analysis	0,59264544
0,47692308	-0,06660232	J48	Physiological	0,4451589
0,47692308	-0,09514371	Bagging	Physiological	0,43154685
0,61538462	0,23240435	Random Forest	Physiological	0,75840267
0,56923077	0	Bayesian Network	Physiological	0,56923077
0,6	0,21831637	Naïve Bayes	Physiological	0,73098982
0,49230769	-0,1283535	SVM	Physiological	0,42911828
0,61538462	0,1370154	Neural Network	Physiological	0,69970179
0,47692308	-0,11503532	J48	Keyboard + Mouse	0,42206008

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,52307692	-0,00298656	Bagging	Keyboard + Mouse	0,52151472
0,52307692	0,03171552	Random Forest	Keyboard + Mouse	0,53966658
0,56923077	0	Bayesian Network	Keyboard + Mouse	0,56923077
0,50769231	0,05368517	Naïve Bayes	Keyboard + Mouse	0,53494785
0,50769231	-0,04	SVM	Keyboard + Mouse	0,48738462
0,29090909	-0,10795455	Neural Network	Keyboard + Mouse	0,25950413
0,50769231	-0,03072349	J48	Keyboard + Sentiment Analysis	0,49209423
0,50769231	-0,04944501	Bagging	Keyboard + Sentiment Analysis	0,48258946
0,6	0,19138756	Random Forest	Keyboard + Sentiment Analysis	0,71483254
0,56923077	0	Bayesian Network	Keyboard + Sentiment Analysis	0,56923077
0,55384615	0,11709602	Naïve Bayes	Keyboard + Sentiment Analysis	0,61869933
0,50769231	-0,08900524	SVM	Keyboard + Sentiment Analysis	0,46250503
0,32727273	0,06091371	Neural Network	Keyboard + Sentiment Analysis	0,34720812
0,43076923	-0,17603912	J48	Keyboard + Physiological	0,35493699
0,49230769	-0,08717689	Bagging	Keyboard + Physiological	0,44938984
0,55384615	0,07005427	Random Forest	Keyboard + Physiological	0,59264544
0,56923077	0	Bayesian Network	Keyboard + Physiological	0,56923077
0,47692308	-0,00545951	Naïve Bayes	Keyboard + Physiological	0,47431931
0,47692308	-0,16807611	SVM	Keyboard + Physiological	0,3967637
0,27272727	-0,05263158	Neural Network	Keyboard + Physiological	0,25837321

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,55384615	0,07823961	J48	Mouse + Sentiment Analysis	0,59717886
0,55384615	0,06172225	Bagging	Mouse + Sentiment Analysis	0,58803078
0,47692308	-0,05741627	Random Forest	Mouse + Sentiment Analysis	0,44953993
0,56923077	0	Bayesian Network	Mouse + Sentiment Analysis	0,56923077
0,49230769	0,04369148	Naïve Bayes	Mouse + Sentiment Analysis	0,51381735
0,58461538	0,14929714	SVM	Mouse + Sentiment Analysis	0,67189679
0,30769231	-0,08695652	Neural Network	Mouse + Sentiment Analysis	0,28093645
0,44615385	-0,12934363	J48	Mouse + Physiological	0,38844669
0,49230769	-0,0971867	Bagging	Mouse + Physiological	0,44446193
0,46153846	-0,08384945	Random Forest	Mouse + Physiological	0,42283871
0,56923077	0	Bayesian Network	Mouse + Physiological	0,56923077
0,50769231	0,05368517	Naïve Bayes	Mouse + Physiological	0,53494785
0,6	0,19829222	SVM	Mouse + Physiological	0,71897533
0,4	0,04231205	Neural Network	Mouse + Physiological	0,41692482
0,47692308	-0,0483871	J48	Sentiment Analysis + Physiological	0,45384615
0,61538462	0,16879795	Bagging	Sentiment Analysis + Physiological	0,71926028
0,58461538	0,15665545	Random Forest	Sentiment Analysis + Physiological	0,67619857
0,6	0,08943966	Bayesian Network	Sentiment Analysis + Physiological	0,65366379

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,6	0,20507996	Naïve Bayes	Sentiment Analysis + Physiological	0,72304798
0,43076923	-0,2189559	SVM	Sentiment Analysis + Physiological	0,33644976
0,55384615	-0,0205739	Neural Network	Sentiment Analysis + Physiological	0,54245138
0,49230769	-0,03075444	J48	Keyboard + Mouse + Sentiment Analysis	0,47716704
0,47692308	-0,09514371	Bagging	Keyboard + Mouse + Sentiment Analysis	0,43154685
0,50769231	0,02163688	Random Forest	Keyboard + Mouse + Sentiment Analysis	0,51867718
0,56923077	0	Bayesian Network	Keyboard + Mouse + Sentiment Analysis	0,56923077
0,52307692	0,07948835	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis	0,56465545
0,55384615	0,06172225	SVM	Keyboard + Mouse + Sentiment Analysis	0,58803078
0,41818182	0,05882353	Neural Network	Keyboard + Mouse + Sentiment Analysis	0,44278075
0,4	-0,25061667	J48	Keyboard + Mouse + Physiological	0,29975333
0,44615385	-0,15956392	Bagging	Keyboard + Mouse + Physiological	0,37496379
0,50769231	-0,003861	Random Forest	Keyboard + Mouse + Physiological	0,50573211

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,56923077	0	Bayesian Network	Keyboard + Mouse + Physiological	0,56923077
0,47692308	-0,02220167	Naïve Bayes	Keyboard + Mouse + Physiological	0,46633459
0,6	0,17721519	SVM	Keyboard + Mouse + Physiological	0,70632911
0,34545455	-0,05263158	Neural Network	Keyboard + Mouse + Physiological	0,32727273
0,46153846	-0,11246944	J48	Keyboard + Sentiment Analysis + Physiological	0,40962949
0,55384615	0,08628211	Bagging	Keyboard + Sentiment Analysis + Physiological	0,60163317
0,6	0,21175373	Random Forest	Keyboard + Sentiment Analysis + Physiological	0,72705224
0,56923077	0	Bayesian Network	Keyboard + Sentiment Analysis + Physiological	0,56923077
0,55384615	0,12447747	Naïve Bayes	Keyboard + Sentiment Analysis + Physiological	0,62278752
0,46153846	-0,1223483	SVM	Keyboard + Sentiment Analysis + Physiological	0,40507002
0,47272727	0,18539326	Neural Network	Keyboard + Sentiment Analysis + Physiological	0,56036772
0,49230769	-0,06769537	J48	Mouse + Sentiment Analysis + Physiological	0,45898074

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,46153846	-0,11246944	Bagging	Mouse + Sentiment Analysis + Physiological	0,40962949
0,50769231	0,01328273	Random Forest	Mouse + Sentiment Analysis + Physiological	0,51443585
0,56923077	0	Bayesian Network	Mouse + Sentiment Analysis + Physiological	0,56923077
0,50769231	0,04587156	Naïve Bayes	Mouse + Sentiment Analysis + Physiological	0,53098095
0,63076923	0,25996205	SVM	Mouse + Sentiment Analysis + Physiological	0,79474529
0,43076923	0,1426025	Neural Network	Mouse + Sentiment Analysis + Physiological	0,492198
0,6	0,18436293	J48	Keyboard + Mouse + Sentiment Analysis + Physiological	0,71061776
0,6	0,155	Bagging	Keyboard + Mouse + Sentiment Analysis + Physiological	0,693
0,52307692	0,04818139	Random Forest	Keyboard + Mouse + Sentiment Analysis + Physiological	0,5482795
0,38461538	-0,34854772	Bayesian Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,25055857

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,50769231	0,03792784	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis + Physiological	0,52694798
0,64615385	0,28159539	SVM	Keyboard + Mouse + Sentiment Analysis + Physiological	0,82810779
0,30434783	-0,03661972	Neural Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,29320269

Table 35. Prediction results for labeling approach 4

Best prediction result per data source

Score	Accuracy	Kappa	Data Source	Algorithm
0,85007692	0,66153846	0,285	Sentiment Analysis	Random Forest
0,82810779	0,64615385	0,28159539	Keyboard + Mouse + Sentiment Analysis + Physiological	SVM
0,79474529	0,63076923	0,25996205	Mouse + Sentiment Analysis + Physiological	SVM
0,75840267	0,61538462	0,23240435	Physiological	Random Forest
0,72705224	0,6	0,21175373	Keyboard + Sentiment Analysis + Physiological	Random Forest
0,72304798	0,6	0,20507996	Sentiment Analysis + Physiological	Naïve Bayes
0,71897533	0,6	0,19829222	Mouse + Physiological	SVM
0,71483254	0,6	0,19138756	Keyboard + Sentiment Analysis	Random Forest
0,70632911	0,6	0,17721519	Keyboard + Mouse + Physiological	SVM
0,67189679	0,58461538	0,14929714	Mouse + Sentiment Analysis	SVM
0,59264544	0,55384615	0,07005427	Keyboard + Physiological	Random Forest
0,58803078	0,55384615	0,06172225	Keyboard + Mouse + Sentiment Analysis	SVM
0,56923077	0,56923077	0	Keyboard	Bayesian Network
0,56923077	0,56923077	0	Keyboard	Bayesian Network

0,56923077	0,56923077	0	Keyboard	Bayesian Network

Table 36. Best prediction per data source for labeling approach 4

13.3.5. Approach 5: Mean SAM arousal values given by participants during the problems in each task.

Prediction results

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,70114943	-0,02260398	J48	Keyboard	0,68530066
0,68965517	-0,04446421	Bagging	Keyboard	0,6589902
0,62068966	-0,03758583	Random Forest	Keyboard	0,59736052
0,71264368	0	Bayesian Network	Keyboard	0,71264368
0,6091954	-0,17942584	Naïve Bayes	Keyboard	0,49989001
0,70114943	-0,02260398	SVM	Keyboard	0,68530066
0,71052632	-0,05025126	Neural Network	Keyboard	0,67482148
0,67816092	0,08283133	J48	Mouse	0,73433389
0,68965517	-0,01119242	Bagging	Mouse	0,68193626
0,68965517	0,21411843	Random Forest	Mouse	0,83732306
0,71264368	0	Bayesian Network	Mouse	0,71264368
0,49425287	-0,08013544	Naïve Bayes	Mouse	0,4546457
0,71264368	0	SVM	Mouse	0,71264368
0,67816092	-0,0656168	Neural Network	Mouse	0,63366217
0,67816092	-0,0656168	J48	Sentiment Analysis	0,63366217
0,71264368	0	Bagging	Sentiment Analysis	0,71264368
0,62068966	0,01509434	Random Forest	Sentiment Analysis	0,63005856
0,71264368	0	Bayesian Network	Sentiment Analysis	0,71264368
0,71264368	0	Naïve Bayes	Sentiment Analysis	0,71264368
0,71264368	0	SVM	Sentiment Analysis	0,71264368
0,63218391	-0,04819277	Neural Network	Sentiment Analysis	0,60171721
0,68965517	0,12773858	J48	Physiological	0,77775075
0,71264368	0,11978956	Bagging	Physiological	0,79801095
0,65517241	0,09312022	Random Forest	Physiological	0,71618221
0,71264368	0	Bayesian Network	Physiological	0,71264368
0,63218391	0,12287335	Naïve Bayes	Physiological	0,70986246
0,72413793	0,05605787	SVM	Physiological	0,76473156
0,70114943	-0,02260398	Neural Network	Physiological	0,68530066
0,70114943	0,04152542	J48	Keyboard + Mouse	0,73026495

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,66666667	-0,08609557	Bagging	Keyboard + Mouse	0,60926962
0,66666667	0,11193242	Random Forest	Keyboard + Mouse	0,74128828
0,71264368	0	Bayesian Network	Keyboard + Mouse	0,71264368
0,54022989	-0,24108417	Naïve Bayes	Keyboard + Mouse	0,40998901
0,70114943	-0,02260398	SVM	Keyboard + Mouse	0,68530066
0,72368421	-0,02570694	Neural Network	Keyboard + Mouse	0,7050805
0,67816092	-0,0656168	J48	Keyboard + Sentiment Analysis	0,63366217
0,68965517	-0,01119242	Bagging	Keyboard + Sentiment Analysis	0,68193626
0,70114943	0,12393493	Random Forest	Keyboard + Sentiment Analysis	0,78804633
0,71264368	0	Bayesian Network	Keyboard + Sentiment Analysis	0,71264368
0,63218391	-0,14379622	Naïve Bayes	Keyboard + Sentiment Analysis	0,54127825
0,70114943	-0,02260398	SVM	Keyboard + Sentiment Analysis	0,68530066
0,73684211	0	Neural Network	Keyboard + Sentiment Analysis	0,73684211
0,65517241	0,01731928	J48	Keyboard + Physiological	0,66651953
0,71264368	0,06371072	Bagging	Keyboard + Physiological	0,75804672
0,6091954	-0,11370482	Random Forest	Keyboard + Physiological	0,53992695
0,71264368	0	Bayesian Network	Keyboard + Physiological	0,71264368
0,65517241	-0,04066986	Naïve Bayes	Keyboard + Physiological	0,62852665
0,72413793	0,11525424	SVM	Keyboard + Physiological	0,8075979
0,72368421	-0,02570694	Neural Network	Keyboard + Physiological	0,7050805

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,66666667	0,03665521	J48	Mouse + Sentiment Analysis	0,69110347
0,71264368	0,03290351	Bagging	Mouse + Sentiment Analysis	0,73609216
0,67816092	0,10769231	Random Forest	Mouse + Sentiment Analysis	0,75119363
0,71264368	0	Bayesian Network	Mouse + Sentiment Analysis	0,71264368
0,49425287	-0,12720848	Naïve Bayes	Mouse + Sentiment Analysis	0,43137972
0,71264368	0	SVM	Mouse + Sentiment Analysis	0,71264368
0,67816092	-0,0656168	Neural Network	Mouse + Sentiment Analysis	0,63366217
0,62068966	0,1260274	J48	Mouse + Physiological	0,69891356
0,70114943	0,01049869	Bagging	Mouse + Physiological	0,70851057
0,63218391	-0,01978022	Random Forest	Mouse + Physiological	0,61967917
0,71264368	0	Bayesian Network	Mouse + Physiological	0,71264368
0,5862069	0,15167931	Naïve Bayes	Mouse + Physiological	0,67512235
0,72413793	0,08661417	SVM	Mouse + Physiological	0,78685854
0,66666667	-0,08609557	Neural Network	Mouse + Physiological	0,60926962
0,68965517	0,12773858	J48	Sentiment Analysis + Physiological	0,77775075
0,71264368	0,06371072	Bagging	Sentiment Analysis + Physiological	0,75804672
0,67816092	0,13124108	Random Forest	Sentiment Analysis + Physiological	0,76716349
0,71264368	0	Bayesian Network	Sentiment Analysis + Physiological	0,71264368

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,59770115	0,05169729	Naïve Bayes	Sentiment Analysis + Physiological	0,62860068
0,72413793	0,05605787	SVM	Sentiment Analysis + Physiological	0,76473156
0,71264368	0	Neural Network	Sentiment Analysis + Physiological	0,71264368
0,64367816	-0,05972495	J48	Keyboard + Mouse + Sentiment Analysis	0,60523451
0,71264368	0,06371072	Bagging	Keyboard + Mouse + Sentiment Analysis	0,75804672
0,59770115	-0,10046982	Random Forest	Keyboard + Mouse + Sentiment Analysis	0,53765022
0,71264368	0	Bayesian Network	Keyboard + Mouse + Sentiment Analysis	0,71264368
0,56321839	-0,17902996	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis	0,46238543
0,70114943	-0,02260398	SVM	Keyboard + Mouse + Sentiment Analysis	0,68530066
0,72368421	-0,02570694	Neural Network	Keyboard + Mouse + Sentiment Analysis	0,7050805
0,64367816	0,1194907	J48	Keyboard + Mouse + Physiological	0,72059171
0,66666667	-0,05256571	Bagging	Keyboard + Mouse + Physiological	0,63162286
0,66666667	0,03665521	Random Forest	Keyboard + Mouse + Physiological	0,69110347

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,71264368	0	Bayesian Network	Keyboard + Mouse + Physiological	0,71264368
0,63218391	0,05691057	Naïve Bayes	Keyboard + Mouse + Physiological	0,66816185
0,67816092	-0,03220339	SVM	Keyboard + Mouse + Physiological	0,65632184
0,72368421	-0,02570694	Neural Network	Keyboard + Mouse + Physiological	0,7050805
0,56321839	-0,09253139	J48	Keyboard + Sentiment Analysis + Physiological	0,51110301
0,67816092	-8,22E+11	Bagging	Keyboard + Sentiment Analysis + Physiological	-5,5724E+11
0,64367816	0,05068638	Random Forest	Keyboard + Sentiment Analysis + Physiological	0,67630388
0,68965517	-0,04446421	Bayesian Network	Keyboard + Sentiment Analysis + Physiological	0,6589902
0,65517241	-0,1059322	Naïve Bayes	Keyboard + Sentiment Analysis + Physiological	0,58576856
0,71264368	0,06371072	SVM	Keyboard + Sentiment Analysis + Physiological	0,75804672
0,72368421	-0,02570694	Neural Network	Keyboard + Sentiment Analysis + Physiological	0,7050805
0,66666667	0,13447684	J48	Mouse + Sentiment Analysis + Physiological	0,7563179

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,68965517	0,07701375	Bagging	Mouse + Sentiment Analysis + Physiological	0,74276811
0,66666667	0,06312662	Random Forest	Mouse + Sentiment Analysis + Physiological	0,70875108
0,71264368	0	Bayesian Network	Mouse + Sentiment Analysis + Physiological	0,71264368
0,54022989	-0,00288184	Naïve Bayes	Mouse + Sentiment Analysis + Physiological	0,53867303
0,72413793	0,11525424	SVM	Mouse + Sentiment Analysis + Physiological	0,8075979
0,72413793	0,05605787	Neural Network	Mouse + Sentiment Analysis + Physiological	0,76473156
0,65517241	0,06918688	J48	Keyboard + Mouse + Sentiment Analysis + Physiological	0,70050175
0,74712644	0,18898305	Bagging	Keyboard + Mouse + Sentiment Analysis + Physiological	0,88832067
0,71264368	0,1453831	Random Forest	Keyboard + Mouse + Sentiment Analysis + Physiological	0,81625003
0,71264368	0	Bayesian Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,71264368

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,62068966	-0,01055966	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis + Physiological	0,61413538
0,72413793	0,14215283	SVM	Keyboard + Mouse + Sentiment Analysis + Physiological	0,82707619
0,69354839	-0,03152364	Neural Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,67168522

Table 37. Prediction results for labeling approach 5

Best prediction result per data source

Score	Accuracy	Kappa	Data Source	Algorithm
0,88832067	0,74712644	0,18898305	Keyboard + Mouse + Sentiment Analysis + Physiological	Bagging
0,83732306	0,68965517	0,21411843	Mouse	Random Forest
0,8075979	0,72413793	0,11525424	Keyboard + Physiological	SVM
0,8075979	0,72413793	0,11525424	Keyboard + Physiological	SVM
0,79801095	0,71264368	0,11978956	Physiological	Bagging
0,78804633	0,70114943	0,12393493	Keyboard + Sentiment Analysis	Random Forest
0,78685854	0,72413793	0,08661417	Mouse + Physiological	SVM
0,77775075	0,68965517	0,12773858	Sentiment Analysis + Physiological	J48
0,75804672	0,71264368	0,06371072	Keyboard + Mouse + Sentiment Analysis	Bagging
0,75804672	0,71264368	0,06371072	Keyboard + Mouse + Sentiment Analysis	Bagging
0,75119363	0,67816092	0,10769231	Mouse + Sentiment Analysis	Random Forest
0,74128828	0,66666667	0,11193242	Keyboard + Mouse	Random Forest
0,72059171	0,64367816	0,1194907	Keyboard + Mouse + Physiological	J48
0,71264368	0,71264368	0	Keyboard	Bayesian Network
0,71264368	0,71264368	0	Keyboard	Bayesian Network

Table 38. Best prediction per data source for labeling approach 5

13.3.6. Approach 6: average of the valence labels presented in the points 2 and 4 in this list.

Prediction results

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,59574468	0,18447489	J48	Keyboard	0,70564461
0,57446809	0,14233577	Bagging	Keyboard	0,65623544
0,5106383	0,0181653	Random Forest	Keyboard	0,5199142
0,61702128	0,22242647	Bayesian Network	Keyboard	0,75426314
0,53191489	0,0582878	Naïve Bayes	Keyboard	0,56291904
0,55319149	0,09366391	SVM	Keyboard	0,60500557
0,28205128	-0,01675978	Neural Network	Keyboard	0,27732417
0,53191489	0,06846847	J48	Mouse	0,56833429
0,44680851	-0,10488246	Bagging	Mouse	0,39994614
0,55319149	0,10678733	Random Forest	Mouse	0,61226533
0,36170213	-0,28181818	Bayesian Network	Mouse	0,25976789
0,59574468	0,20125224	Naïve Bayes	Mouse	0,71563963
0,53191489	0,05656934	SVM	Mouse	0,56200497
0,36170213	-0,02322206	Neural Network	Mouse	0,35330266
0,80851064	0,61580381	J48	Sentiment Analysis	1,30639457
0,80851064	0,61580381	Bagging	Sentiment Analysis	1,30639457
0,68085106	0,36429216	Random Forest	Sentiment Analysis	0,92887977
0,80851064	0,61580381	Bayesian Network	Sentiment Analysis	1,30639457
0,74468085	0,48913043	Naïve Bayes	Sentiment Analysis	1,10892692
0,78723404	0,57272727	SVM	Sentiment Analysis	1,23810445
0,76595745	0,53127833	Neural Network	Sentiment Analysis	1,17289404
0,59574468	0,18744313	J48	Physiological	0,70741293
0,53191489	0,06509946	Bagging	Physiological	0,56654226
0,70212766	0,40181818	Random Forest	Physiological	0,98425532
0,46808511	-0,08494922	Bayesian Network	Physiological	0,42832164
0,5106383	0,01278539	Naïve Bayes	Physiological	0,51716701
0,46808511	-0,06915378	SVM	Physiological	0,43571525
0,46808511	-0,07699358	Neural Network	Physiological	0,43204556
0,61702128	0,23646209	J48	Keyboard + Mouse	0,76292342

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,55319149	0,11000902	Bagging	Keyboard + Mouse	0,61404754
0,59574468	0,19038985	Random Forest	Keyboard + Mouse	0,70916842
0,57446809	0,13602941	Bayesian Network	Keyboard + Mouse	0,65261264
0,59574468	0,20409982	Naïve Bayes	Keyboard + Mouse	0,71733606
0,59574468	0,18596171	SVM	Keyboard + Mouse	0,70653038
0,20512821	0,00247525	Neural Network	Keyboard + Mouse	0,20563595
0,80851064	0,61719457	J48	Keyboard + Sentiment Analysis	1,30751901
0,80851064	0,61580381	Bagging	Keyboard + Sentiment Analysis	1,30639457
0,82978723	0,65880218	Random Forest	Keyboard + Sentiment Analysis	1,37645287
0,82978723	0,65942029	Bayesian Network	Keyboard + Sentiment Analysis	1,37696577
0,59574468	0,18891916	Naïve Bayes	Keyboard + Sentiment Analysis	0,70829227
0,68085106	0,35967302	SVM	Keyboard + Sentiment Analysis	0,92573483
0,46153846	-0,06640625	Neural Network	Keyboard + Sentiment Analysis	0,43088942
0,57446809	0,143898	J48	Keyboard + Physiological	0,65713289
0,46808511	-0,06334842	Bagging	Keyboard + Physiological	0,43843266
0,55319149	0,10516772	Random Forest	Keyboard + Physiological	0,61136938
0,59574468	0,17998163	Bayesian Network	Keyboard + Physiological	0,70296778
0,5106383	0,01637853	Naïve Bayes	Keyboard + Physiological	0,5190018
0,4893617	-0,03296703	SVM	Keyboard + Physiological	0,4732289
0,41025641	0,03548387	Neural Network	Keyboard + Physiological	0,4248139

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,72340426	0,44404004	J48	Mouse + Sentiment Analysis	1,04462471
0,70212766	0,40506329	Bagging	Mouse + Sentiment Analysis	0,9865338
0,72340426	0,44504995	Random Forest	Mouse + Sentiment Analysis	1,04535529
0,76595745	0,53042688	Bayesian Network	Mouse + Sentiment Analysis	1,17224187
0,61702128	0,22669104	Naïve Bayes	Mouse + Sentiment Analysis	0,75689447
0,70212766	0,39963504	SVM	Mouse + Sentiment Analysis	0,98272247
0,31914894	-0,00133156	Neural Network	Mouse + Sentiment Analysis	0,31872397
0,42553191	-0,15468608	J48	Mouse + Physiological	0,35970805
0,46808511	-0,06721163	Bagging	Mouse + Physiological	0,43662435
0,59574468	0,19038985	Random Forest	Mouse + Physiological	0,70916842
0,27659574	-0,4580292	Bayesian Network	Mouse + Physiological	0,14990682
0,57446809	0,143898	Naïve Bayes	Mouse + Physiological	0,65713289
0,55319149	0,09698079	SVM	Mouse + Physiological	0,60684044
0,40425532	0,04775687	Neural Network	Mouse + Physiological	0,42356129
0,63829787	0,27953111	J48	Sentiment Analysis + Physiological	0,81672198
0,76595745	0,53127833	Bagging	Sentiment Analysis + Physiological	1,17289404
0,74468085	0,48727273	Random Forest	Sentiment Analysis + Physiological	1,10754352
0,80851064	0,61580381	Bayesian Network	Sentiment Analysis + Physiological	1,30639457

Accuracy	**Coehn's Kappa**	**Predictor**	**Data Sources**	**Score**
0,63829787	0,27561197	Naïve Bayes	Sentiment Analysis + Physiological	0,8142204
0,63829787	0,27429609	SVM	Sentiment Analysis + Physiological	0,81338049
0,38297872	-0,24702653	Neural Network	Sentiment Analysis + Physiological	0,28837282
0,87234043	0,74456522	J48	Keyboard + Mouse + Sentiment Analysis	1,52185476
0,80851064	0,61580381	Bagging	Keyboard + Mouse + Sentiment Analysis	1,30639457
0,80851064	0,61369863	Random Forest	Keyboard + Mouse + Sentiment Analysis	1,30469251
0,82978723	0,65942029	Bayesian Network	Keyboard + Mouse + Sentiment Analysis	1,37696577
0,61702128	0,22810219	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis	0,75776518
0,70212766	0,4007286	SVM	Keyboard + Mouse + Sentiment Analysis	0,98349029
0,53846154	0,17605634	Neural Network	Keyboard + Mouse + Sentiment Analysis	0,63326111
0,57446809	0,14855072	J48	Keyboard + Mouse + Physiological	0,65980574
0,4893617	-0,02173913	Bagging	Keyboard + Mouse + Physiological	0,4787234
0,65957447	0,31884058	Random Forest	Keyboard + Mouse + Physiological	0,86987357

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,44680851	-0,12110092	Bayesian Network	Keyboard + Mouse + Physiological	0,39269959
0,59574468	0,18744313	Naïve Bayes	Keyboard + Mouse + Physiological	0,70741293
0,59574468	0,18596171	SVM	Keyboard + Mouse + Physiological	0,70653038
0,30769231	-0,10377358	Neural Network	Keyboard + Mouse + Physiological	0,27576197
0,78723404	0,57504521	J48	Keyboard + Sentiment Analysis + Physiological	1,23992921
0,76595745	0,53042688	Bagging	Keyboard + Sentiment Analysis + Physiological	1,17224187
0,70212766	0,40181818	Random Forest	Keyboard + Sentiment Analysis + Physiological	0,98425532
0,80851064	0,61650045	Bayesian Network	Keyboard + Sentiment Analysis + Physiological	1,30695781
0,53191489	0,06	Naïve Bayes	Keyboard + Sentiment Analysis + Physiological	0,56382979
0,55319149	0,10354223	SVM	Keyboard + Sentiment Analysis + Physiological	0,61047017
0,35897436	-0,0483871	Neural Network	Keyboard + Sentiment Analysis + Physiological	0,34160463
0,70212766	0,40290381	J48	Mouse + Sentiment Analysis + Physiological	0,98501757

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,76595745	0,53042688	Bagging	Mouse + Sentiment Analysis + Physiological	1,17224187
0,61702128	0,2323049	Random Forest	Mouse + Sentiment Analysis + Physiological	0,76035834
0,80851064	0,61580381	Bayesian Network	Mouse + Sentiment Analysis + Physiological	1,30639457
0,61702128	0,23090909	Naïve Bayes	Mouse + Sentiment Analysis + Physiological	0,7594971
0,63829787	0,27164995	SVM	Mouse + Sentiment Analysis + Physiological	0,81169146
0,36170213	0,06062625	Neural Network	Mouse + Sentiment Analysis + Physiological	0,38363077
0,78723404	0,57350272	J48	Keyboard + Mouse + Sentiment Analysis + Physiological	1,23871491
0,78723404	0,57350272	Bagging	Keyboard + Mouse + Sentiment Analysis + Physiological	1,23871491
0,59574468	0,19038985	Random Forest	Keyboard + Mouse + Sentiment Analysis + Physiological	0,70916842
0,76595745	0,52957234	Bayesian Network	Keyboard + Mouse + Sentiment Analysis + Physiological	1,17158732

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,65957447	0,31636364	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis + Physiological	0,86823985
0,68085106	0,35967302	SVM	Keyboard + Mouse + Sentiment Analysis + Physiological	0,92573483
0,375	-0,01426307	Neural Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,36965135

Table 39. Prediction results for labeling approach 6

Best prediction result per data source

Score	Accuracy	Kappa	Data Source	Algorithm
1,52185476	0,87234043	0,74456522	Keyboard + Mouse + Sentiment Analysis	J48
1,37696577	0,82978723	0,65942029	Keyboard + Sentiment Analysis	Bayesian Network
1,30695781	0,80851064	0,61650045	Keyboard + Sentiment Analysis + Physiological	Bayesian Network
1,30639457	0,80851064	0,61580381	Sentiment Analysis	J48
1,30639457	0,80851064	0,61580381	Sentiment Analysis	J48
1,30639457	0,80851064	0,61580381	Sentiment Analysis	J48
1,23871491	0,78723404	0,57350272	Keyboard + Mouse + Sentiment Analysis + Physiological	J48
1,17224187	0,76595745	0,53042688	Mouse + Sentiment Analysis	Bayesian Network
0,98425532	0,70212766	0,40181818	Physiological	Random Forest
0,86987357	0,65957447	0,31884058	Keyboard + Mouse + Physiological	Random Forest
0,76292342	0,61702128	0,23646209	Keyboard + Mouse	J48
0,75426314	0,61702128	0,22242647	Keyboard	Bayesian Network
0,71563963	0,59574468	0,20125224	Mouse	Naïve Bayes
0,70916842	0,59574468	0,19038985	Mouse + Physiological	Random Forest
0,70296778	0,59574468	0,17998163	Keyboard + Physiological	Bayesian Network

Table 40. Best prediction per data source for labeling approach 6

13.3.7. Approach 7: average of the arousal labels presented in the points 3 and 5 in this list.

Prediction results

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,65217391	0,03664921	J48	Keyboard	0,67607557
0,58695652	-0,08436725	Bagging	Keyboard	0,53743662
0,63043478	0,17336152	Random Forest	Keyboard	0,73972792
0,65217391	0	Bayesian Network	Keyboard	0,65217391
0,52173913	-0,19339623	Naïve Bayes	Keyboard	0,42083675
0,63043478	-0,04266667	SVM	Keyboard	0,60353623
0,65217391	0,03664921	Neural Network	Keyboard	0,67607557
0,54347826	-0,19851117	J48	Mouse	0,43559176
0,58695652	-0,12339332	Bagging	Mouse	0,51453001
0,63043478	0,09280742	Random Forest	Mouse	0,68894381
0,65217391	0	Bayesian Network	Mouse	0,65217391
0,36956522	-0,33133733	Naïve Bayes	Mouse	0,24711447
0,60869565	-0,08376963	SVM	Mouse	0,55770544
0,56521739	-0,16161616	Neural Network	Mouse	0,47386913
0,56521739	-0,08490566	J48	Sentiment Analysis	0,51722724
0,63043478	-0,04266667	Bagging	Sentiment Analysis	0,60353623
0,63043478	0,21956088	Random Forest	Sentiment Analysis	0,7688536
0,65217391	0	Bayesian Network	Sentiment Analysis	0,65217391
0,60869565	-0,08376963	Naïve Bayes	Sentiment Analysis	0,55770544
0,65217391	0	SVM	Sentiment Analysis	0,65217391
0,60869565	0,0840708	Neural Network	Sentiment Analysis	0,65986918
0,56521739	0,04166667	J48	Physiological	0,58876812
0,69565217	0,18686869	Bagging	Physiological	0,82564778
0,76086957	0,39328537	Random Forest	Physiological	1,06010843
0,60869565	-0,08376963	Bayesian Network	Physiological	0,55770544
0,60869565	0,11158798	Naïve Bayes	Physiological	0,67661877
0,7173913	0,25806452	SVM	Physiological	0,90252454
0,65217391	0	Neural Network	Physiological	0,65217391
0,56521739	-0,08490566	J48	Keyboard + Mouse	0,51722724

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,58695652	-0,12339332	Bagging	Keyboard + Mouse	0,51453001
0,69565217	0,21463415	Random Forest	Keyboard + Mouse	0,84496288
0,58695652	-0,12339332	Bayesian Network	Keyboard + Mouse	0,51453001
0,45652174	-0,21564482	Naïve Bayes	Keyboard + Mouse	0,35807519
0,60869565	-0,08376963	SVM	Keyboard + Mouse	0,55770544
0,67391304	0,08	Neural Network	Keyboard + Mouse	0,72782609
0,63043478	0,06235012	J48	Keyboard + Sentiment Analysis	0,66974247
0,60869565	-0,08376963	Bagging	Keyboard + Sentiment Analysis	0,55770544
0,63043478	0,14814815	Random Forest	Keyboard + Sentiment Analysis	0,72383253
0,65217391	0	Bayesian Network	Keyboard + Sentiment Analysis	0,65217391
0,56521739	-0,16161616	Naïve Bayes	Keyboard + Sentiment Analysis	0,47386913
0,63043478	-0,04266667	SVM	Keyboard + Sentiment Analysis	0,60353623
0,65217391	0	Neural Network	Keyboard + Sentiment Analysis	0,65217391
0,7173913	0,38603696	J48	Keyboard + Physiological	0,99433086
0,63043478	-0,00514139	Bagging	Keyboard + Physiological	0,62719347
0,7173913	0,32808989	Random Forest	Keyboard + Physiological	0,95276014
0,63043478	-0,00514139	Bayesian Network	Keyboard + Physiological	0,62719347
0,54347826	-0,05228758	Naïve Bayes	Keyboard + Physiological	0,5150611
0,63043478	0,02977667	SVM	Keyboard + Physiological	0,64920703
0,63043478	-0,04266667	Neural Network	Keyboard + Physiological	0,60353623

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,52173913	-0,11946903	J48	Mouse + Sentiment Analysis	0,45940746
0,54347826	-0,15827338	Bagging	Mouse + Sentiment Analysis	0,45746012
0,69565217	0,32916667	Random Forest	Mouse + Sentiment Analysis	0,92463768
0,65217391	0	Bayesian Network	Mouse + Sentiment Analysis	0,65217391
0,45652174	-0,21564482	Naïve Bayes	Mouse + Sentiment Analysis	0,35807519
0,65217391	0	SVM	Mouse + Sentiment Analysis	0,65217391
0,58695652	-0,08436725	Neural Network	Mouse + Sentiment Analysis	0,53743662
0,76086957	0,44880174	J48	Mouse + Physiological	1,10234915
0,60869565	-0,08376963	Bagging	Mouse + Physiological	0,55770544
0,67391304	0,17266187	Random Forest	Mouse + Physiological	0,79027213
0,65217391	0,03664921	Bayesian Network	Mouse + Physiological	0,67607557
0,56521739	0,11877395	Naïve Bayes	Mouse + Physiological	0,63235049
0,67391304	0,17266187	SVM	Mouse + Physiological	0,79027213
0,58695652	-0,12339332	Neural Network	Mouse + Physiological	0,51453001
0,73913043	0,425	J48	Sentiment Analysis + Physiological	1,05326087
0,58695652	-0,08436725	Bagging	Sentiment Analysis + Physiological	0,53743662
0,73913043	0,3490566	Random Forest	Sentiment Analysis + Physiological	0,99712879
0,60869565	-0,08376963	Bayesian Network	Sentiment Analysis + Physiological	0,55770544

Accuracy	**Coehn's Kappa**	**Predictor**	**Data Sources**	**Score**
0,63043478	0,19712526	Naïve Bayes	Sentiment Analysis + Physiological	0,7547094
0,67391304	0,1439206	SVM	Sentiment Analysis + Physiological	0,77090301
0,65217391	0	Neural Network	Sentiment Analysis + Physiological	0,65217391
0,56521739	0,04166667	J48	Keyboard + Mouse + Sentiment Analysis	0,58876812
0,63043478	-0,00514139	Bagging	Keyboard + Mouse + Sentiment Analysis	0,62719347
0,52173913	-0,11946903	Random Forest	Keyboard + Mouse + Sentiment Analysis	0,45940746
0,65217391	0	Bayesian Network	Keyboard + Mouse + Sentiment Analysis	0,65217391
0,36956522	-0,36960986	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis	0,23297027
0,60869565	-0,08376963	SVM	Keyboard + Mouse + Sentiment Analysis	0,55770544
0,60869565	-0,04545455	Neural Network	Keyboard + Mouse + Sentiment Analysis	0,58102767
0,7826087	0,52083333	J48	Keyboard + Mouse + Physiological	1,19021739
0,58695652	-0,04796163	Bagging	Keyboard + Mouse + Physiological	0,55880513
0,69565217	0,21463415	Random Forest	Keyboard + Mouse + Physiological	0,84496288

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,60869565	-0,08376963	Bayesian Network	Keyboard + Mouse + Physiological	0,55770544
0,54347826	-0,05228758	Naïve Bayes	Keyboard + Mouse + Physiological	0,5150611
0,56521739	-0,08490566	SVM	Keyboard + Mouse + Physiological	0,51722724
0,58695652	-0,12339332	Neural Network	Keyboard + Mouse + Physiological	0,51453001
0,7173913	0,38603696	J48	Keyboard + Sentiment Analysis + Physiological	0,99433086
0,65217391	0,07070707	Bagging	Keyboard + Sentiment Analysis + Physiological	0,69828722
0,69565217	0,18686869	Random Forest	Keyboard + Sentiment Analysis + Physiological	0,82564778
0,63043478	-0,04266667	Bayesian Network	Keyboard + Sentiment Analysis + Physiological	0,60353623
0,60869565	0,11158798	Naïve Bayes	Keyboard + Sentiment Analysis + Physiological	0,67661877
0,63043478	0,09280742	SVM	Keyboard + Sentiment Analysis + Physiological	0,68894381
0,63043478	-0,04266667	Neural Network	Keyboard + Sentiment Analysis + Physiological	0,60353623
0,7826087	0,53441296	J48	Mouse + Sentiment Analysis + Physiological	1,20084492

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,58695652	-0,12339332	Bagging	Mouse + Sentiment Analysis + Physiological	0,51453001
0,76086957	0,41299304	Random Forest	Mouse + Sentiment Analysis + Physiological	1,0751034
0,65217391	0,03664921	Bayesian Network	Mouse + Sentiment Analysis + Physiological	0,67607557
0,60869565	0,16194332	Naïve Bayes	Mouse + Sentiment Analysis + Physiological	0,70726985
0,7173913	0,3062645	SVM	Mouse + Sentiment Analysis + Physiological	0,93710279
0,60869565	-0,08376963	Neural Network	Mouse + Sentiment Analysis + Physiological	0,55770544
0,76086957	0,43146067	J48	Keyboard + Mouse + Sentiment Analysis + Physiological	1,08915486
0,7173913	0,25806452	Bagging	Keyboard + Mouse + Sentiment Analysis + Physiological	0,90252454
0,80434783	0,54901961	Random Forest	Keyboard + Mouse + Sentiment Analysis + Physiological	1,24595055
0,67391304	0,24836601	Bayesian Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,84129014

Accuracy	Coehn's Kappa	Predictor	Data Sources	Score
0,63043478	0,12134831	Naïve Bayes	Keyboard + Mouse + Sentiment Analysis + Physiological	0,70693698
0,67391304	0,19953596	SVM	Keyboard + Mouse + Sentiment Analysis + Physiological	0,80838293
0,46153846	-0,24279211	Neural Network	Keyboard + Mouse + Sentiment Analysis + Physiological	0,34948056

Table 41. Prediction results for labeling approach 7

Best prediction result per data source

Score	Accuracy	Kappa	Data Source	Algorithm
1,24595055	0,80434783	0,54901961	Keyboard + Mouse + Sentiment Analysis + Physiological	Random Forest
1,20084492	0,7826087	0,53441296	Mouse + Sentiment Analysis + Physiological	J48
1,19021739	0,7826087	0,52083333	Keyboard + Mouse + Physiological	J48
1,10234915	0,76086957	0,44880174	Mouse + Physiological	J48
1,06010843	0,76086957	0,39328537	Physiological	Random Forest
1,05326087	0,73913043	0,425	Sentiment Analysis + Physiological	J48
0,99433086	0,7173913	0,38603696	Keyboard + Physiological	J48
0,99433086	0,7173913	0,38603696	Keyboard + Physiological	J48
0,92463768	0,69565217	0,32916667	Mouse + Sentiment Analysis	Random Forest
0,84496288	0,69565217	0,21463415	Keyboard + Mouse	Random Forest
0,7688536	0,63043478	0,21956088	Sentiment Analysis	Random Forest
0,73972792	0,63043478	0,17336152	Keyboard	Random Forest
0,72383253	0,63043478	0,14814815	Keyboard + Sentiment Analysis	Random Forest
0,68894381	0,63043478	0,09280742	Mouse	Random Forest
0,65217391	0,65217391	0	Keyboard + Mouse + Sentiment Analysis	Bayesian Network

Table 42. Best prediction per data source for labeling approach 7

13.4. Appendix IV : Full results from the data mining processing in transition stage

This section reports the full results obtained in the data mining process for models generated in the transition stage.

The following table depicts the model generation process and the accuracy, Cohen's Kappa values and the score proposed in formula (4.6).

Prediction results

Model generation process	Accuracy	Kohen's Kappa	Score
2-step classification:J48	0,74796748	0,49433062	1,1177107
2-step classification:PCA->Bagging	0,70731707	0,41308238	0,99949729
2-step classification:Bagging	0,69105691	0,37981822	0,95353292
2-step classification:BFE(NB)&NB	0,6504065	0,29933099	0,84509333
2-step classification:NaiveBayes	0,64634146	0,28263056	0,82901731
2-step classification:BFE(Bagging)&Bagging	0,61788618	0,23130111	0,76080394
2-step classification:BFE(J48)&J48	0,58943089	0,17864463	0,69472956
2-step classification:PCA->J48	0,57723577	0,14634635	0,66171212
2-step classification:PCA->NaiveBayes	0,56504065	0,10911799	0,62669675
2-step classification:BFE(NB)&J48	0,55691057	0,10934697	0,61780706
2-step classification:Bagging->NonesRemoved->PCA->Bagging	0,6300813	0,34822105	0,84948887
2-step classification:Bagging->NonesRemoved->PCA->J48	0,62601626	0,35136994	0,84597956
2-step classification:J48->NonesRemoved->BFE(NB)&J48	0,62601626	0,33086953	0,83314596
2-step classification:J48->NonesRemoved->Bagging	0,61788618	0,34052019	0,8282889
PCA->J48->NonesRemoved->PCA->Bagging	0,62601626	0,30629885	0,81776432
2-step classification:J48->NonesRemoved->BFE(Bagging)&Bagging	0,62195122	0,30489472	0,81158086
BFE(Bagging)&Bagging->NonesRemoved->PCA->Bagging	0,61788618	0,30147414	0,80416289
2-step classification:Bagging->NonesRemoved->BFE(J48)&J48	0,61382114	0,29891402	0,79730088
2-step classification:J48->NonesRemoved->BFE(J48)&J48	0,6097561	0,29716378	0,79095352
2-step classification:J48->NonesRemoved->PCA->Bagging	0,60162602	0,3127904	0,78980886
2-step classification:J48->NonesRemoved->BFE(NB)&NB	0,59756098	0,3203851	0,78901061

Model generation process	Accuracy	Kohen's Kappa	Score
PCA->Bagging->NonesRemoved->PCA->Bagging	0,60162602	0,30504468	0,78514883
2-step classification:Bagging->NonesRemoved->Bagging	0,60162602	0,29683535	0,78020989
PCA->Bagging	0,59349594	0,31156074	0,77840597
PCA->Bagging->NonesRemoved->BFE(J48)&J48	0,60569106	0,28307896	0,77714945
BFE(NB)&NB->NonesRemoved->PCA->Bagging	0,60162602	0,28426803	0,77264906
BFE(Bagging)&Bagging->NonesRemoved->PCA->J48	0,60162602	0,28294816	0,77185499
2-step classification:Bagging->NonesRemoved->BFE(Bagging)&Bagging	0,60162602	0,27677447	0,76814074
BFE(NB)&NB->NonesRemoved->BFE(J48)&J48	0,60162602	0,27339582	0,76610806
BFE(J48)&J48->NonesRemoved->PCA->J48	0,59349594	0,28105912	0,76030338
2-step classification:Bagging->NonesRemoved->BFE(NB)&J48	0,59349594	0,26963957	0,75352592
PCA->J48->NonesRemoved->PCA->J48	0,59756098	0,25768105	0,75154112
PCA->Bagging->NonesRemoved->PCA->J48	0,58130081	0,29217532	0,75114257
BFE(J48)&J48->NonesRemoved->BFE(J48)&J48	0,59349594	0,26457399	0,75051952
BFE(J48)&J48->NonesRemoved->PCA->Bagging	0,59349594	0,26378165	0,75004927
BFE(Bagging)&Bagging->NonesRemoved->BFE(NB)&J48	0,58943089	0,26612713	0,74629445
2-step classification:NaiveBayes->NonesRemoved->PCA->J48	0,59349594	0,2571118	0,74609074
2-step classification:NaiveBayes->NonesRemoved->PCA->Bagging	0,59349594	0,25089071	0,74239855
2-step classification:Bagging->NonesRemoved->BFE(NB)&NB	0,58130081	0,26467003	0,73515372
2-step classification:NaiveBayes->NonesRemoved->Bagging	0,58943089	0,24475652	0,73369795
2-step classification:NaiveBayes->NonesRemoved->BFE(NB)&J48	0,58943089	0,23465993	0,72774671
2-step classification:J48->NonesRemoved->PCA->J48	0,56910569	0,27847261	0,72758604
PCA->Bagging->NonesRemoved->BFE(NB)&J48	0,58130081	0,24911095	0,72610921

Model generation process	**Accuracy**	**Kohen's Kappa**	**Score**
PCA->Bagging->NonesRemoved->BFE(Bagging)&Bagging	0,57723577	0,2533489	0,72347782
PCA->J48->NonesRemoved->BFE(J48)&J48	0,58536585	0,23523316	0,72306331
2-step classification:NaiveBayes->NonesRemoved->BFE(J48)&J48	0,58943089	0,22636692	0,72285855
BFE(J48)&J48->NonesRemoved->BFE(Bagging)&Bagging	0,58130081	0,24312214	0,72262791
BFE(NB)&NB->NonesRemoved->BFE(NB)&J48	0,57317073	0,25713957	0,72055561
PCA->J48	0,56097561	0,27757233	0,71668692
2-step classification:NaiveBayes->NonesRemoved->BFE(Bagging)&Bagging	0,57723577	0,21890456	0,70359532
PCA->J48->NonesRemoved->BFE(NB)&NB	0,57317073	0,22504575	0,70216037
BFE(NB)&J48->NonesRemoved->BFE(J48)&J48	0,57723577	0,21192706	0,69956765
Class Missing Values Removed->PCA->J48	0,56302521	0,23971004	0,69798801
BFE(Bagging)&Bagging->NonesRemoved->BFE(Bagging)&Bagging	0,57317073	0,2156565	0,69677873
BFE(NB)&J48->NonesRemoved->BFE(NB)&J48	0,57317073	0,21417706	0,69593076
BFE(NB)&NB->NonesRemoved->BFE(Bagging)&Bagging	0,56910569	0,2157594	0,69189559
BFE(Bagging)&Bagging->NonesRemoved->BFE(J48)&J48	0,56504065	0,21581362	0,68698412
BFE(Bagging)&Bagging->NonesRemoved->BFE(NB)&NB	0,56097561	0,21317301	0,68056047
Class Missing Values Removed->BFE(NB)&J48	0,59663866	0,13755096	0,67870688
BFE(NB)&J48->NonesRemoved->PCA->Bagging	0,56504065	0,18976822	0,67226741
BFE(NB)&J48->NonesRemoved->PCA->J48	0,56504065	0,18846925	0,67153344
BFE(J48)&J48->NonesRemoved->BFE(NB)&J48	0,56097561	0,19707456	0,67152963
BFE(J48)&J48	0,56097561	0,18742354	0,66611564
PCA->Bagging->NonesRemoved->BFE(NB)&NB	0,54065041	0,22658728	0,66315491
BFE(NB)&NB->NonesRemoved->PCA->J48	0,54471545	0,21383325	0,66119372

Model generation process	**Accuracy**	**Kohen's Kappa**	**Score**
PCA->J48->NonesRemoved->BFE(NB)&J48	0,55691057	0,18710968	0,66111393
2-step classification:NaiveBayes->NonesRemoved->BFE(NB)&NB	0,55284553	0,19401918	0,66010817
Class Missing Values Removed->PCA->Bagging	0,55462185	0,18228964	0,65572367
PCA->J48->NonesRemoved->BFE(Bagging)&Bagging	0,55284553	0,18252674	0,65375462
2-step classification:Bagging->NonesRemoved->J48	0,51376147	0,25448445	0,64450577
BFE(NB)&J48->NonesRemoved->BFE(Bagging)&Bagging	0,54878049	0,17219426	0,64327734
Class Missing Values Removed->BFE(NB)&NB	0,53781513	0,18462688	0,63711026
BFE(NB)&J48	0,54471545	0,15549425	0,62941557
Class Missing Values Removed->BFE(Bagging)&Bagging	0,57142857	0,09888641	0,62793509
BFE(J48)&J48->NonesRemoved->BFE(NB)&NB	0,51626016	0,19828011	0,61862429
2-step classification:J48->NonesRemoved->PCA->NaiveBayes	0,50406504	0,2245362	0,61724589
BFE(Bagging)&Bagging->NonesRemoved->PCA->NaiveBayes	0,51626016	0,15632025	0,59696208
BFE(NB)&NB	0,49593496	0,20060798	0,59542347
PCA->NaiveBayes->NonesRemoved->BFE(J48)&J48	0,54065041	0,09857967	0,59394755
2-step classification:NaiveBayes->NonesRemoved->PCA->NaiveBayes	0,52439024	0,1301641	0,59264703
BFE(NB)&NB->NonesRemoved->BFE(NB)&NB	0,49593496	0,19063918	0,59047959
PCA->NaiveBayes->NonesRemoved->PCA->Bagging	0,53658537	0,09916161	0,58979403
BFE(NB)&NB->NonesRemoved->PCA->NaiveBayes	0,52439024	0,12370224	0,58925849
2-step classification:NaiveBayes->NonesRemoved->J48	0,5	0,17033525	0,58516763
BFE(Bagging)&Bagging	0,52845529	0,105903	0,58442028
Class Missing Values Removed->BFE(J48)&J48	0,54621849	0,06923523	0,58403605
BFE(NB)&J48->NonesRemoved->BFE(NB)&NB	0,51219512	0,13214758	0,57988047
PCA->NaiveBayes->NonesRemoved->BFE(NB)&NB	0,52845529	0,09253959	0,57735832

Model generation process	Accuracy	Kohen's Kappa	Score
PCA->NaiveBayes->NonesRemoved->BFE(Bagging)&Bagging	0,53252033	0,08268483	0,57655168
PCA->NaiveBayes->NonesRemoved->PCA->NaiveBayes	0,53252033	0,08035888	0,57531306
2-step classification:J48->NonesRemoved->J48	0,46846847	0,22193181	0,57243652
PCA->NaiveBayes->NonesRemoved->BFE(NB)&J48	0,52845529	0,08217812	0,57188275
PCA->J48->NonesRemoved->PCA->NaiveBayes	0,50813008	0,12026009	0,56923785
PCA->NaiveBayes->NonesRemoved->PCA->J48	0,52439024	0,06778947	0,55993838
BFE(NB)&J48->NonesRemoved->PCA->NaiveBayes	0,51219512	0,08448083	0,55546579
PCA->Bagging->NonesRemoved->PCA->NaiveBayes	0,48373984	0,13789012	0,55044278
2-step classification:Bagging->NonesRemoved->PCA->NaiveBayes	0,47154472	0,14288012	0,53891908
BFE(J48)&J48->NonesRemoved->PCA->NaiveBayes	0,43089431	0,09575446	0,47215436
2-step classification:J48->NonesRemoved->NaiveBayes	0,36036036	0,20905259	0,43569463
2-step classification:Bagging->NonesRemoved->NaiveBayes	0,35779817	0,1631019	0,41615573
2-step classification:NaiveBayes->NonesRemoved->NaiveBayes	0,34285714	0,062318	0,36422331
Class Missing Values Removed->PCA->NaiveBayes	0,24369748	0,08998216	0,2656259
PCA->NaiveBayes	0,17479675	0,00255663	0,17524364

Table 43. Results from the transition stage predictions

13.5. **Appendix V : Full results from the data mining processing in stage 2**

In this section top 100 models (according to the score calculated using formula (4.6)) for valence and for arousal prediction are to be presented.

13.5.1. Top valence prediction models

Score	**Accuracy**	**Cohen's kappa**	**Accuracy improvement**	**Algorithm**	**FFS+PCA**	**ESS+SMOTE**	**numBinsTargetAtt**	**physioBLname**	**useBL**	**Clustering**	**2-step classification**
0,310	0,717	0,433	0,21	SMO	FFS	ESS	2	LBpost	Dynamic	None	No
0,296	0,708	0,417	0,20	BayesNet	FFS	SMOTE	2	LBpost	Dynamic	None	No
0,296	0,708	0,417	0,20	RF	Raw	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,296	0,708	0,417	0,20	RF	Raw	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,278	0,700	0,397	0,19	BayesNet	Raw	SMOTE	2	LBpre	Dynamic	None	No
0,265	0,692	0,383	0,18	Bagging	FFS	ESS	2	copiaTexto	Dynamic	None	No
0,264	0,692	0,382	0,18	SMO	FFS	Raw	2	LBpre	Dynamic	None	No
0,251	0,683	0,368	0,18	RF	FFS	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,251	0,683	0,368	0,18	RF	Both	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,251	0,683	0,367	0,18	RF	Raw	SMOTE	2	LBpost	Dynamic	EMclustering	No+clustering
0,251	0,683	0,367	0,18	RF	PCA	SMOTE	2	LBpost	Dynamic	EMclustering	No+clustering
0,251	0,683	0,367	0,18	RF	PCA	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,251	0,683	0,367	0,18	RF	Raw	SMOTE	2	LBpost	Dynamic	EMclustering	No+clustering
0,249	0,683	0,365	0,18	RF	FFS	Raw	2	copiaTexto	Dynamic	None	No
0,249	0,683	0,364	0,18	BayesNet	Raw	SMOTE	2	LBpre	Dynamic	None	No
0,238	0,675	0,353	0,17	J48	Both	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,237	0,675	0,351	0,17	Bagging	FFS	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,237	0,675	0,351	0,17	Bagging	Raw	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,237	0,675	0,351	0,17	Bagging	Raw	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,236	0,675	0,350	0,17	Bagging	PCA	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,236	0,675	0,350	0,17	BayesNet	FFS	Both	2	LBcombi	Dynamic	None	No
0,236	0,675	0,350	0,17	SMO	Both	ESS	2	copiaTexto	Dynamic	None	No
0,236	0,675	0,350	0,17	BayesNet	FFS	Both	2	LBpre	Dynamic	None	No
0,236	0,675	0,349	0,17	SMO	FFS	Raw	2	LBcombi	Dynamic	None	No
0,235	0,675	0,349	0,17	BayesNet	FFS	SMOTE	2	LBpre	Dynamic	None	No
0,235	0,675	0,348	0,17	BayesN	Raw	SMOT	2	LBcom	Dynam	None	No

				et		E		bi	ic		
0,234	0,675	0,347	0,17	BayesNet	Raw	SMOTE	2	LBcombi	Dynamic	None	No
0,225	0,650	0,347	0,05	RF	PCA	Raw	2	LBpre	NoBL	EMclustering	No+clustering
0,225	0,650	0,347	0,05	RF	PCA	Raw	2	copiaTexto	NoBL	EMclustering	No+clustering
0,225	0,650	0,347	0,05	RF	PCA	Raw	2	LBpost	NoBL	EMclustering	No+clustering
0,225	0,650	0,347	0,05	RF	PCA	Raw	2	LBcombi	NoBL	EMclustering	No+clustering
0,222	0,667	0,334	0,16	Bagging	FFS	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,222	0,667	0,333	0,16	Bagging	Both	ESS	2	LBcombi	Dynamic	None	No
0,221	0,667	0,332	0,16	BayesNet	FFS	Both	2	LBpost	Dynamic	None	No
0,221	0,667	0,331	0,16	SMO	FFS	Raw	2	LBpost	Dynamic	None	No
0,220	0,667	0,331	0,16	SMO	FFS	SMOTE	2	LBpre	Dynamic	None	No
0,220	0,667	0,330	0,16	SMO	FFS	SMOTE	2	LBpost	Dynamic	None	No
0,213	0,594	0,359	0,14	RF	FFS	Both	3	copiaTexto	NoBL	CascadeSimpleKMeans	No+clustering
0,213	0,619	0,344	0,16	BayesNet	FFS	Raw	3	copiaTexto	NoBL	None	No
0,212	0,606	0,350	0,15	SMO	Raw	Raw	3	LBcombi	NoBL	None	Yes
0,212	0,606	0,350	0,15	SMO	Raw	Raw	3	copiaTexto	NoBL	None	Yes
0,212	0,606	0,350	0,15	SMO	Raw	Raw	3	LBpre	NoBL	None	Yes
0,212	0,606	0,350	0,15	SMO	Raw	Raw	3	LBpost	NoBL	None	Yes
0,212	0,606	0,350	0,15	SMO	Raw	Raw	3	LBcombi	NoBL	None	Yes
0,212	0,606	0,350	0,15	SMO	Raw	Raw	3	copiaTexto	NoBL	None	Yes
0,212	0,606	0,350	0,15	SMO	Raw	Raw	3	LBpre	NoBL	None	Yes
0,212	0,606	0,350	0,15	SMO	Raw	Raw	3	LBpost	NoBL	None	Yes
0,210	0,663	0,317	0,06	Bagging	PCA	SMOTE	2	LBpre	NoBL	EMclustering	No+clustering
0,210	0,663	0,317	0,06	Bagging	PCA	SMOTE	2	LBpost	NoBL	EMclustering	No+clustering
0,210	0,663	0,317	0,06	Bagging	PCA	SMOTE	2	copiaTexto	NoBL	EMclustering	No+clustering
0,210	0,663	0,317	0,06	Bagging	PCA	SMOTE	2	LBcombi	NoBL	EMclustering	No+clustering
0,210	0,658	0,319	0,15	Bagging	FFS	SMOTE	2	LBpre	Dynamic	EMclustering	No+clustering
0,209	0,658	0,318	0,15	Bagging	Both	SMOTE	2	LBpost	Dynamic	EMclustering	No+clustering
0,209	0,658	0,318	0,15	J48	Raw	ESS	2	LBcombi	Dynamic	None	No
0,209	0,658	0,317	0,15	Bagging	FFS	SMOTE	2	LBpost	Dynamic	EMclustering	No+clustering
0,208	0,658	0,316	0,15	Bagging	Both	Both	2	LBcombi	Dynamic	EMclustering	No+clustering
0,208	0,658	0,316	0,15	Bagging	PCA	Both	2	LBcombi	Dynamic	EMclustering	No+clustering
0,208	0,658	0,316	0,15	BayesNet	FFS	SMOTE	2	LBpost	Dynamic	None	No

0,208	0,658	0,316	0,15	BayesNet	FFS	Both	2	LBcombi	Dynamic	None	No
0,208	0,658	0,316	0,15	BayesNet	FFS	SMOTE	2	copiaTexto	Dynamic	None	No
0,208	0,658	0,316	0,15	RF	FFS	Raw	2	LBpost	Dynamic	None	No
0,207	0,600	0,346	0,14	RF	Both	Raw	3	LBcombi	NoBL	EMclusteringtwoStepClassification	Yes+clustering
0,207	0,658	0,315	0,15	SMO	FFS	Raw	2	copiaTexto	Dynamic	None	No
0,207	0,658	0,315	0,15	BayesNet	FFS	Both	2	copiaTexto	Dynamic	None	No
0,207	0,658	0,315	0,15	SMO	FFS	Both	2	LBpre	Dynamic	None	No
0,207	0,658	0,315	0,15	BayesNet	Raw	SMOTE	2	LBpost	Dynamic	None	No
0,207	0,658	0,315	0,15	BayesNet	FFS	SMOTE	2	LBpre	Dynamic	None	No
0,207	0,658	0,314	0,15	BayesNet	Raw	SMOTE	2	LBpost	Dynamic	None	No
0,207	0,658	0,314	0,15	BayesNet	Raw	Raw	2	LBpre	Dynamic	None	No
0,206	0,658	0,314	0,15	SMO	FFS	Raw	2	LBcombi	Dynamic	None	No
0,201	0,588	0,341	0,13	Bagging	Raw	SMOTE	3	LBcombi	NoBL	None	Yes
0,201	0,588	0,341	0,13	Bagging	Raw	SMOTE	3	copiaTexto	NoBL	None	Yes
0,201	0,588	0,341	0,13	Bagging	Raw	SMOTE	3	LBpre	NoBL	None	Yes
0,201	0,588	0,341	0,13	Bagging	Raw	SMOTE	3	LBpost	NoBL	None	Yes
0,200	0,594	0,336	0,14	J48	FFS	Raw	3	LBpost	NoBL	EMclustering	No+clustering
0,199	0,631	0,315	0,03	RF	Raw	Raw	2	copiaTexto	NoBL	EMclustering	No+clustering
0,199	0,631	0,315	0,03	RF	Raw	Raw	2	LBcombi	NoBL	EMclustering	No+clustering
0,199	0,631	0,315	0,03	RF	Raw	Raw	2	LBpre	NoBL	EMclustering	No+clustering
0,199	0,631	0,315	0,03	RF	Raw	Raw	2	LBpost	NoBL	EMclustering	No+clustering
0,199	0,631	0,315	0,03	RF	Raw	Raw	2	copiaTexto	NoBL	EMclustering	No+clustering
0,199	0,631	0,315	0,03	RF	Raw	Raw	2	LBcombi	NoBL	EMclustering	No+clustering
0,199	0,631	0,315	0,03	RF	Raw	Raw	2	LBpre	NoBL	EMclustering	No+clustering
0,199	0,631	0,315	0,03	RF	Raw	Raw	2	LBpost	NoBL	EMclustering	No+clustering
0,199	0,575	0,346	0,12	J48	Raw	Raw	3	LBcombi	NoBL	None	Yes
0,199	0,575	0,346	0,12	J48	Raw	Raw	3	copiaTexto	NoBL	None	Yes
0,199	0,575	0,346	0,12	J48	Raw	Raw	3	LBpre	NoBL	None	Yes
0,199	0,575	0,346	0,12	J48	Raw	Raw	3	LBpost	NoBL	None	Yes
0,199	0,581	0,342	0,13	RF	Both	Both	3	LBpost	NoBL	CascadeSimpleKMeans	No+clustering
0,197	0,575	0,343	0,12	BayesNet	Raw	SMOTE	3	LBcombi	NoBL	None	Yes

0,197	0,575	0,343	0,12	BayesNet	Raw	SMOTE	3	copiaTexto	NoBL	None	Yes
0,197	0,575	0,343	0,12	BayesNet	Raw	SMOTE	3	LBpre	NoBL	None	Yes
0,197	0,575	0,343	0,12	BayesNet	Raw	SMOTE	3	LBpost	NoBL	None	Yes
0,196	0,650	0,301	0,14	Bagging	FFS	SMOTE	2	copiaTexto	Dynamic	EMclustering	No+clustering
0,196	0,650	0,301	0,14	RF	Both	SMOTE	2	LBpost	Dynamic	EMclustering	No+clustering
0,196	0,650	0,301	0,14	Bagging	PCA	SMOTE	2	LBpost	Dynamic	EMclustering	No+clustering
0,196	0,650	0,301	0,14	RF	FFS	Both	2	LBpost	Dynamic	EMclustering	No+clustering
0,195	0,708	0,276	0,04	Bagging	Raw	SMOTE	2	LBcombi	Fixed	None	No
0,195	0,650	0,301	0,14	RF	FFS	SMOTE	2	LBpost	Dynamic	EMclustering	No+clustering
0,195	0,650	0,301	0,14	Bagging	Both	SMOTE	2	LBcombi	Dynamic	EMclustering	No+clustering
0,195	0,650	0,300	0,14	Bagging	PCA	Both	2	copiaTexto	Dynamic	EMclustering	No+clustering

Table 44. Best 100 results in stage 2 predicting valence

13.5.2. Top arousal prediction models

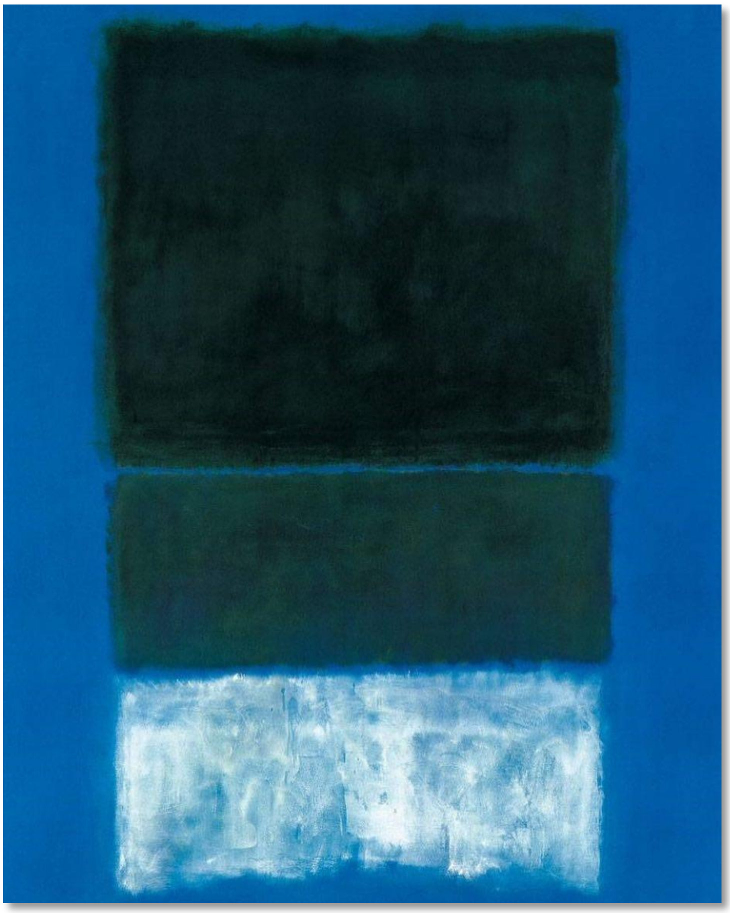
Score	Accuracy	Cohen's kappa	Accuracy improvement	Algorithm	FFS+PCA	ESS+SMOTE	numBinsTargetAtt	physioBLname	useBL	Clustering	2-step classification
0,301	0,675	0,445	0,20625	BayesNet	FFS	SMOTE	3	LBpre	NoBL	None	No
0,291	0,675	0,430	0,20625	SMO	FFS	Raw	3	LBpre	NoBL	None	No
0,287	0,669	0,429	0,2	BayesNet	Both	Raw	3	LBcombi	NoBL	None	No
0,285	0,669	0,426	0,2	SMO	Raw	Raw	3	LBcombi	NoBL	None	No
0,285	0,669	0,426	0,2	SMO	Raw	Raw	3	copiaTexto	NoBL	None	No
0,285	0,669	0,426	0,2	SMO	Raw	Raw	3	LBpre	NoBL	None	No
0,285	0,669	0,426	0,2	SMO	Raw	Raw	3	LBpost	NoBL	None	No
0,282	0,663	0,426	0,19375	SMO	FFS	Raw	3	copiaTexto	NoBL	None	Yes
0,280	0,669	0,419	0,2	SMO	FFS	Raw	3	LBpost	NoBL	None	No
0,278	0,663	0,419	0,19375	BayesNet	FFS	SMOTE	3	LBcombi	NoBL	None	No
0,276	0,656	0,421	0,1875	SMO	Raw	Raw	3	LBcombi	NoBL	None	Yes
0,276	0,656	0,421	0,1875	SMO	Raw	Raw	3	copiaTexto	NoBL	None	Yes
0,276	0,656	0,421	0,1875	SMO	Raw	Raw	3	LBpre	NoBL	None	Yes
0,276	0,656	0,421	0,1875	SMO	Raw	Raw	3	LBpost	NoBL	None	Yes
0,274	0,656	0,417	0,1875	BayesNet	FFS	SMOTE	3	copiaTexto	NoBL	None	No
0,271	0,656	0,413	0,1875	BayesNet	FFS	Raw	3	copiaTexto	NoBL	None	No
0,271	0,656	0,413	0,1875	BayesNet	FFS	Raw	3	LBpost	NoBL	None	No
0,271	0,656	0,413	0,1875	BayesNet	FFS	Raw	3	LBpre	NoBL	None	No
0,271	0,656	0,413	0,1875	BayesNet	FFS	Raw	3	LBcombi	NoBL	None	No
0,270	0,656	0,411	0,1875	Bagging	FFS	Raw	3	LBpost	NoBL	None	No
0,270	0,631	0,427	0,1625	J48	FFS	ESS	3	LBcombi	NoBL	None	No

0,266	0,656	0,405	0,1875	J48	FFS	Raw	3	LBpost	NoBL	None	No
0,265	0,656	0,403	0,1875	Bagging	FFS	Raw	3	LBcombi	NoBL	None	No
0,264	0,708	0,373	-0,025	SMO	Raw	Both	2	copiaTexto	Dynamic	None	No
0,263	0,656	0,401	0,1875	Bagging	Both	Raw	3	copiaTexto	NoBL	None	No
0,261	0,644	0,405	0,175	J48	FFS	Raw	3	LBpost	NoBL	None	Yes
0,259	0,650	0,398	0,18125	Bagging	FFS	Raw	3	LBpre	NoBL	None	No
0,256	0,650	0,394	0,18125	BayesNet	Both	Raw	3	LBpost	NoBL	None	No
0,255	0,644	0,397	0,175	SMO	FFS	Raw	3	LBpre	NoBL	None	Yes
0,255	0,644	0,396	0,175	SMO	FFS	SMOTE	3	LBpost	NoBL	None	Yes
0,255	0,758	0,336	0,025	BayesNet	Both	SMOTE	2	LBpost	Dynamic	Cascade SimpleKMeans	No+clustering
0,254	0,644	0,395	0,175	SMO	FFS	SMOTE	3	copiaTexto	NoBL	None	Yes
0,253	0,650	0,390	0,18125	SMO	FFS	Raw	3	LBcombi	NoBL	None	No
0,253	0,783	0,323	-0,01666667	J48	FFS	SMOTE	2	LBpre	Fixed	None	No
0,252	0,650	0,388	0,18125	SMO	FFS	SMOTE	3	copiaTexto	NoBL	None	No
0,251	0,650	0,387	0,18125	SMO	FFS	SMOTE	3	LBcombi	NoBL	None	No
0,251	0,631	0,397	0,1625	NB	FFS	Raw	3	LBcombi	NoBL	None	No
0,249	0,638	0,391	0,16875	Bagging	FFS	Raw	3	LBpost	NoBL	None	Yes
0,248	0,631	0,392	0,1625	RF	Raw	SMOTE	3	LBcombi	NoBL	None	Yes
0,248	0,631	0,392	0,1625	RF	Raw	SMOTE	3	copiaTexto	NoBL	None	Yes
0,248	0,631	0,392	0,1625	RF	Raw	SMOTE	3	LBpre	NoBL	None	Yes
0,248	0,631	0,392	0,1625	RF	Raw	SMOTE	3	LBpost	NoBL	None	Yes
0,247	0,767	0,323	0,03333333	SMO	Raw	SMOTE	2	LBcombi	Dynamic	None	No
0,247	0,767	0,323	0,03333333	Bagging	PCA	SMOTE	2	LBcombi	Dynamic	Cascade SimpleKMeans	No+clustering
0,247	0,638	0,387	0,16875	J48	FFS	Raw	3	LBcombi	NoBL	None	Yes
0,246	0,638	0,385	0,16875	Bagging	FFS	Raw	3	LBpre	NoBL	None	Yes
0,243	0,644	0,378	0,175	SMO	FFS	SMOTE	3	LBpost	NoBL	None	No
0,243	0,638	0,381	0,16875	RF	Both	Raw	3	copiaTexto	NoBL	None	No
0,242	0,625	0,388	0,15625	J48	Both	Raw	3	LBpost	NoBL	None	No
0,242	0,644	0,376	0,175	SMO	FFS	Raw	3	copiaTexto	NoBL	None	No
0,242	0,758	0,319	-0,04166667	NB	Raw	SMOTE	2	LBcombi	Fixed	None	No
0,241	0,817	0,295	0,01666667	Bagging	FFS	SMOTE	2	LBcombi	Fixed	None	No
0,240	0,631	0,380	0,1625	J48	FFS	Raw	3	copiaTexto	NoBL	None	Yes
0,238	0,638	0,374	0,16875	BayesNet	FFS	SMOTE	3	LBpost	NoBL	None	No
0,238	0,625	0,380	0,15625	BayesNet	Both	Raw	3	copiaTexto	NoBL	None	Yes
0,237	0,625	0,380	0,15625	SMO	Raw	SMOTE	3	LBcombi	NoBL	None	Yes
0,237	0,625	0,380	0,15625	SMO	Raw	SMOTE	3	copiaTexto	NoBL	None	Yes
0,237	0,625	0,380	0,15625	SMO	Raw	SMOTE	3	LBpre	NoBL	None	Yes

0,237	0,625	0,380	0,15625	SMO	Raw	SMOTE	3	LBpost	NoBL	None	Yes
0,237	0,625	0,379	0,15625	Bagging	FFS	SMOTE	3	LBcombi	NoBL	None	Yes
0,237	0,638	0,371	0,16875	J48	FFS	Raw	3	LBpre	NoBL	None	No
0,236	0,625	0,377	0,15625	BayesNet	Both	Raw	3	LBpre	NoBL	None	Yes
0,236	0,825	0,286	0,025	RF	FFS	Raw	2	LBcombi	Fixed	Cascade SimpleKMeans	No+clustering
0,235	0,631	0,373	0,1625	J48	FFS	Raw	3	LBcombi	NoBL	None	No
0,235	0,631	0,372	0,1625	J48	FFS	Raw	3	copiaTexto	NoBL	None	No
0,235	0,638	0,368	0,16875	SMO	FFS	SMOTE	3	LBpre	NoBL	None	No
0,234	0,638	0,368	0,16875	SMO	FFS	SMOTE	3	copiaTexto	NoBL	None	No
0,234	0,625	0,374	0,15625	Bagging	FFS	Raw	3	copiaTexto	NoBL	None	Yes
0,234	0,631	0,370	0,1625	SMO	FFS	Raw	3	LBcombi	NoBL	None	Yes
0,234	0,631	0,370	0,1625	SMO	FFS	SMOTE	3	LBcombi	NoBL	None	Yes
0,233	0,631	0,369	0,1625	SMO	Both	Raw	3	copiaTexto	NoBL	None	Yes
0,233	0,675	0,345	-0,05833333	RF	FFS	ESS	2	copiaTexto	Dynamic	None	No
0,232	0,758	0,306	0,025	NB	FFS	SMOTE	2	LBpost	Dynamic	Cascade SimpleKMeans	No+clustering
0,232	0,619	0,375	0,15	RF	FFS	SMOTE	3	LBcombi	NoBL	None	Yes
0,229	0,638	0,360	0,16875	SMO	FFS	SMOTE	3	LBpost	NoBL	None	No
0,229	0,625	0,367	0,15625	SMO	FFS	Raw	3	LBpost	NoBL	None	Yes
0,229	0,619	0,370	0,15	Bagging	Both	Raw	3	LBpre	NoBL	None	Yes
0,228	0,783	0,291	0,05	SMO	Raw	Raw	2	LBcombi	Dynamic	None	No
0,227	0,625	0,363	0,15625	Bagging	Both	Raw	3	LBpost	NoBL	None	No
0,226	0,638	0,355	0,16875	SMO	Both	Raw	3	LBcombi	NoBL	None	No
0,226	0,625	0,362	0,15625	Bagging	Raw	Raw	3	copiaTexto	NoBL	None	Yes
0,226	0,625	0,362	0,15625	Bagging	Raw	Raw	3	LBpre	NoBL	None	Yes
0,226	0,625	0,362	0,15625	Bagging	Raw	Raw	3	LBpost	NoBL	None	Yes
0,226	0,625	0,362	0,15625	Bagging	Raw	Raw	3	LBcombi	NoBL	None	Yes
0,226	0,619	0,365	0,15	J48	FFS	Raw	3	LBpre	NoBL	None	Yes
0,226	0,625	0,361	0,15625	SMO	FFS	SMOTE	3	LBpre	NoBL	None	Yes
0,225	0,619	0,363	0,15	Bagging	FFS	Raw	3	LBcombi	NoBL	None	Yes
0,223	0,767	0,291	0,03333333	NB	Both	Raw	2	LBcombi	Dynamic	None	No
0,222	0,683	0,325	-0,05	SMO	FFS	Both	2	copiaTexto	Dynamic	None	No
0,222	0,631	0,351	0,1625	SMO	Both	Raw	3	LBpre	NoBL	None	No
0,221	0,613	0,361	0,14375	J48	Both	Raw	3	LBpost	NoBL	None	Yes
0,220	0,700	0,315	-0,03333333	J48	FFS	ESS	2	copiaTexto	Dynamic	None	No
0,220	0,758	0,290	0,025	NB	PCA	Raw	2	LBcombi	Dynamic	Cascade SimpleKMeans	No+clustering
0,219	0,631	0,347	0,1625	SMO	Both	Raw	3	copiaTexto	NoBL	None	No
0,218	0,606	0,359	0,1375	RF	FFS	SMOTE	3	copiaTe	NoBL	None	Yes

								xto			
0,218	0,750	0,290	0,01666667	SMO	Raw	SMOTE	2	copiaTexto	Dynamic	None	No
0,218	0,750	0,290	0,01666667	BayesNet	PCA	SMOTE	2	LBpost	Dynamic	CascadeSimpleKMeans	No+clustering
0,217	0,619	0,351	0,15	SMO	Raw	SMOTE	3	LBcombi	NoBL	None	No
0,217	0,619	0,351	0,15	SMO	Raw	SMOTE	3	copiaTexto	NoBL	None	No
0,217	0,619	0,351	0,15	SMO	Raw	SMOTE	3	LBpre	NoBL	None	No

Table 45. Best 100 results in stage 2 predicting arousal



No. 14 (White and Greens in Blue)

Mark Rothko