Using pre-trained language models to automatically identify research phases in biomedical publications

Master's Thesis

Nicolau Duran Silva

Master's Degree in Language Technologies

Universidad Nacional de Educación a Distancia

Supervisors

Prof. Dr. Laura Plaza Morales

Prof. Dr. Jorge Carrillo-de-Albornoz Cuadrado

Advisory supervisor

Dr. Francesco Alessandro Massucci

June 2022

Acknowledgement

Firstly, I would like to thank my supervisors, Jorge Carrillo-de-Albornoz and Laura Plaza, for their support and guidance during these months which have been indispensable for the success of this Master's Thesis.

Special thanks also to Francesco Massucci for his back-up and confidence.

Sara Ricardo and Sonia Veiga for their total involvement and invaluable help in defining the task and dataset. Also, special acknowledgement to Arnau Ramos, who has been unconditionally involved in the project.

I would like to thank SIRIS Academic for its commitment to R&D projects such as this one, and the team for the good atmosphere and inspiring environment.

And for their attention to detail, Adrià Plazas, Alessandro Seri, and Mike Pritchard.

Finally, I must express my sincerest thanks to my parents, and specially to Joana for their constant support and help, and for bearing the brunt.

Resumen

La ciencia, la investigación y la innovación buscan resolver retos complejos, como por ejemplo abordar un tipo de cáncer o, como recientemente, desarrollar la vacuna del COVID-19. La resolución de estos problemas complejos, especialmente en la investigación biomédica, puede ser costosa, ineficiente e insostenible. Suele implicar la colaboración de un amplio conjunto de sectores y actores, puesto que generalmente una sola institución no dispone de los recursos necesarios para desarrollar una innovación de principio a fin, algunos actores se apoyan en otros para combinar sus descubrimientos y lograr una mayor contribución al individuo. De hecho, el número de publicaciones científicas disponibles crece año tras año, especialmente en el ámbito biomédico. Las agencias de financiación, los gobiernos y las universidades están cada vez más interesados en comprender qué actividades de investigación se financian o se llevan a cabo en el ecosistema de investigación, cómo contribuye la ciencia a estas misiones y desafíos, y si existen lagunas de financiación e investigación en diferentes áreas o dominios.

La comprensión de los temas abordados por las publicaciones científicas ha atraído la atención de los investigadores en procesamiento del lenguaje natural (PLN), des de hace varias decadas. Sin embargo, los "dominios específicos", como la biomedicina, se enfrentan a retos y complejidades adicionales. Los modelos neuronales del lenguaje basados en el Transformer han supuesto un gran avance para diversas tareas de PLN, ya que están preentrenados sobre grandes conjuntos de documentos sin etiquetar y son capaces de aprender una representación universal del lenguaje que se adapta a las tareas posteriores. La mayoría de estos modelos están preentrenados sobre textos de dominio general, aunque hay algunos preentrenados o adaptados a los dominios biomédico y clínico, que son especialmente prometedores para abordar el procesamiento y comprensión de textos en el dominio que nos ocupa.

En el presente trabajo, y para dar respuesta a la creciente necesidad de conocer el estado de la investigación en el dominio biomedico, presentamos BATRACIO (*BAsic-TRAnslational-Clinical research phases classification in bIOmedical publications*), un conjunto de datos para clasificar publicaciones científicas del dominio biomédico en fases de investigación. Exploramos si los modelos lingüísticos preentrenados específicos del dominio superan a los modelos del lenguaje preentrenados en el dominio general, y cómo los adaptamos para enfrentarnos a un conjunto de datos desequilibrado en el dominio biomédico y con categorías adyacentes.

Finalmente, en los resultados observamos que los modelos preentrenados del lenguaje basados en BERT, específicamente los modelos preentrenados en el dominio biomédico o científico, ofrecen una gran oportunidad para resolver esta tarea satisfactoriamente. Además, también hemos explorado cómo utilizarlos para la clasificación de textos y qué estrategias pueden ser favorables para la clasificación de artículos de investigación biomédica, como la limpieza del texto y el ajuste de hiperparámetros. No obstante, los principales retos específicos de nuestro conjunto de datos son el desequilibrio de clases y que las categorías no son mutuamente independientes, sino que tienen relaciones semánticas de adyacencia entre ellas. Este no era un objetivo principal del proyecto, pero también hemos explorado si ligeras modificaciones en la función de pérdida pueden hacer frente a las categorías desequilibradas y adyacentes, aunque los resultados de estos experimentos son parcialmente satisfactorios, apuntan a futuras líneas de investigación.

Abstract

Science, research, and innovation, aim to solve complex challenges, such as tackling a specific type of cancer or, as recently, the vaccine for COVID-19. Solving these complex problems, especially in biomedical research, can be expensive, inefficient and unsustainable. It involves collaboration from a broad set of actors, because a complete discovery often requires the involvement of many actors and a single institution does not usually have the resources to develop an innovation from beginning to end, and some actors rely on others to combine their discoveries to achieve a greater contribution to the individual. Indeed, the number of scientific publications available is growing year by year, especially in the biomedical domain. Funding agencies, governments, and universities, are more and more interested in understanding what research activities are funded or carried out in the research and innovation ecosystem, how science is contributing to these missions and challenges, or whether there are funding gaps in different areas or domains.

Understanding topics addressed by scientific publications have attracted attention from researchers in NLP. However, "specific domains" such as biomedicine, face additional challenges and complexity. Transformer-based neural language models, like BERT, have led to breakthroughs for a variety of natural language processing (NLP) tasks, which are pre-trained on large-scale unlabelled documents and can learn universal language representation which is adapted to downstream tasks. Most of these models are pre-trained on general domain data, although there are some which are pre-trained or adapted to the biomedical and clinical domains, which are especially promising for addressing domain texts.

In this Master's Thesis, we present BATRACIO (*BAsic-TRAnslational-Clinical research phases classification in bIOmedical publications*), a dataset for classifying scientific publications in biomedical domain in research phases. We explore if domain specific pre-trained language models outperform gen-

eral pre-trained language models, and how we adapt them to face an imbalanced dataset in biomedical domain with adjacent categories.

Finally, we have seen in results that state-of-the-art BERT-based pre-trained language models, specifically pre-trained in the biomedical or scientific domain, offer a great opportunity to solve this task. Furthermore, we have also explored how to use them for text classification and which strategies may be favourable for the classification of biomedical research articles, such as text cleaning and hyperparameter setting. Nevertheless, the main specific challenges of our dataset are the class imbalance and that categories are not mutually independent, they have semantic relations of adjacency between them. This was not a main goal of the project, but we have also explored whether slight modifications in the loss function can deal with imbalanced and adjacent categories, although the results of these experiments are partially satisfactory, they point to future lines of research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The number of scientific publications produced increases from year to year, but, especially in recent years, this growth has been more significant. The increase in the literature available is most pronounced in the biomedical domain, where more than 3,000 publications appear every day (Kim et al., 2018).

Understanding the topics addressed by scientific publications is a problem that has been attracting attention for more than twenty years (Mcllwaine and Williamson, 1999; Crawford-Welch and McCleary, 1992). During the last few years, several initiatives have attempted to simplify the complexity and to extract knowledge from scientific outputs through the use of language technologies on their textual content. These technologies range from automatic text summarisation (Kieuvongngam et al., 2020), named entity recognition (Giorgi and Bader, 2019), relation extraction (Wei et al., 2020) (Tran et al., 2021), automatic question-answering (Sarrouti and Alaoui, 2020) or automatic text classification (Resnik et al., 2020; Su et al., 2020; You et al., 2020). Text is an extremely rich source of information, although extracting insights or knowledge from it can be time-consuming and enormously challenging because of its unstructured nature (Minaee et al., 2021).

Natural language processing (NLP) aims to study human language so that computers can understand natural language as humans do, and its applications range from identifying parts of speech of words to answering questions, to name a few. In contrast to "general domain" natural language

processing, “specialised domains” like biomedicine face additional problems, and past work has shown that using in-domain text can provide additional advantages over general-domain strategies (Gu et al., 2021; Cohen, 2014). The application of NLP in the biomedical and clinical domains is known as BioNLP, and in the last twenty years research on clinical texts and biomedical literature has grown exponentially.

Some of the main difficulties in this domain comes from lexical ambiguities, ranging from different meanings of different words in different contexts to the disambiguation of acronyms and abbreviations (Duque et al., 2018). For instance, the acronym ”BSA” in the title *The interaction between BSA and DOTAP at the air-buffer interface* can refer to multiple concepts such as “Bovine Serum Albuminum” or “Body Surface Area” (Duque et al., 2018). In addition to word sense disambiguation, other challenges that BioNLP has to deal with are detection of negations, temporal link detection and context determination (Cohen, 2014).

Automatic text classification is a classic problem in natural language processing, which is defined as the task of labelling natural language texts with a set of predefined tags (Sebastiani, 2001), or assigning categories to text units such as sentences, paragraphs or documents (Minaee et al., 2021). This task is very important in many natural language processing applications (Li et al., 2020), such as: Sentiment Analysis, News Categorization, Topic Analysis, Spam Detection, Question Answering (QA), to mention just some of them. Textual data to be classified may come from multiple sources, such as websites, emails, social media, video transcripts, user reviews, tickets, among others (Minaee et al., 2021). In the biomedical domain, text classification tasks have focused on automatically assigning documents with categories related to drugs (Li et al., 2017), diseases (Yao et al., 2018a), genes (Su et al., 2020), and Medical Subject Headings (MeSH) terms (You et al., 2020), among others. These tasks face many additional challenges, since medical terminology is quite specific and contains many abbreviations and acronyms of domain-specific concepts (Qing et al., 2019).

The approaches and techniques used to carry out automatic text classification have evolved over the years, but the arrival of deep learning, neural network-based methods, has changed the NLP game rules in very few years, and some of the most notable developments have been around pre-trained language models (PLM) (Qiu et al., 2020). Pre-trained language models are

trained on large-scale unlabelled corpora of documents and can learn universal language representations based on contextual embeddings. Their main advantage is that they can be applied in a downstream task, avoiding completely training a new model from scratch and overfitting on small datasets, because most available datasets for most supervised NLP tasks are small, due to the extreme expensive annotation costs (Qiu et al., 2020). Probably one of the most relevant breakthroughs in NLP of recent years is the arrival of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), that proposes a bidirectional multilayer Transformer encoder based architecture, which uses bidirectional self-attention technique, where each token can attend to the context on the right and left. Specifically, BERT has performed state-of-the-art results on 11 NLP tasks, and BERT-based models pre-trained on scientific and biomedical literature, BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), or PubMedBERT (Gu et al., 2022), have achieved state-of-the-art results in BioNLP tasks such as named entity recording, relation extraction and question-answering. In this context, BERT-based models have received special attention from the research community for classification of scientific literature, and also for a wide range of tasks in the biomedical domain.

Despite these significant advances in language representation and text classification techniques, the need for high quality and large annotated corpora with the labels to be assigned remains in order to carry out automatic text classification tasks. Annotating a dataset is costly and time-consuming. However, in the biomedical domain it is even more complex, since each task has to be very well-defined and annotators have to be domain experts, due to the need for expertise and knowledge about the concepts and type of documents. For example, in the case of indexing publications in the MEDLINE database with Medical Subject Headings (MeSH), the cost of manually annotating each scientific publication indexed in the database by an expert is estimated in $9.4 (Mork et al., 2013; You et al., 2020). There is no need to calculate how much it would cost to annotate a dataset with ten thousand documents.

A major constraint of classification systems for scientific literature is that they consider categories as independent of each other, even considering the possibility of labelling each document with more than one category, but in practice not all areas are independent of each other. In fact, science tends

to join and become increasingly interdisciplinary. In many cases, areas are specially interconnected, and the development of scientific discoveries or innovation is not done only by a single activity (Todeva and Rakhmatullin, 2016). Science, research and innovation, are about solving complex challenges, such as tackling a specific type of cancer or, as recently, the vaccine for COVID-19. Solving these problems involves collaboration from a broad set of sectors and actors, and requires an extensive proposal of solutions that aim to respond to the challenge, by taking risks, doing experiments, and, in some cases, doing some successful trials (Mazzucato, 2021). This is because a complete discovery often requires the involvement of many actors doing different things in a world where research organizations are getting more and more specialised, where a single institution does not have the resources to develop an innovation from beginning to end, and where some actors rely on others to combine their discoveries to achieve a greater contribution to the individual (Talmar et al., 2020). Solving some of these complex challenges, especially from biomedical research, can be expensive, inefficient and unsustainable (Mazzucato, 2021).

We have found that research is very complex, and it is difficult to understand what research activities are carried out in practice, especially when complex value chains are involved. However, funding agencies, governments, universities and researchers are interested in understanding in which areas they are most active, specialised or collaborate internationally (Fuster et al., 2020). One way to analyse research portfolios to understand research capabilities is to start by exploring research outputs, such as scientific publications (Fuster et al., 2020). Classifying research outputs in research areas that involve complex phases can be really challenging, but it can be extremely useful for a wide range of stakeholders. For instance, it can benefit research funding agencies in understanding what research activities they are funding; government, in mapping the actors researching in a specific area of interest in their territory; for universities, understanding in which areas they are active in practice or to better priority-setting for their research strategy; or for researchers, to most easily extract information from large collections of scientific articles.

In biomedical research, we can find a value chain, composed of the phases of research, which is the focus of this thesis. In a similar manner

as described in (Simpkin et al., 2017; Head et al., 2014) [1] and according to our work done with domain experts, a classification of biomedical research by phases of research should contain the following four categories: `basic research`, `translational research`, `clinical research`, and `public health`. There is a growing interest in classifying biomedical research in research phases, and we have identified a gap in both the definition of these categories and in the automatic classification of documents in them.

This master's thesis explores automatic text classification of scientific publications in the biomedical domain in research phases. We have tackled the problem of classification on the basis of these four categories, defining the borders and formalising the problem, resulting in a framework for defining each phase in detail, even in ambiguous cases. The project defines the BATRACIO task (the acronym of *BAsic-TRAnslational-Clinical research phases classification in bIOmedical publications*), a new BioNLP task. To address the problem, a corpus of biomedical research texts annotated with the categories has been developed, in conjunction with the description of how to create an annotated dataset for this domain-intensive task.

We have also explored how state-of-the-art BERT-based pre-trained language models perform to solve this task. Different language models pre-trained from the open domain to the specific domain have been explored, and the results have been compared with other machine-learning techniques, to assess whether the powerful BERT, and specifically with pre-trained models in the biomedical and scientific domains, obtains better results and whether a better encoding of the context improves the solving of the task proposed, how to use them for text classification and which strategies may be favourable for the classification of biomedical research articles. To the best of the author's knowledge, no one before has addressed the problem of automatically classifying scientific articles in the biomedical research phases. This project is a collaboration with the company SIRIS Academic, and experts in the biomedical field from SIRIS Academic and from the Medical University of Innsbruck have been involved in order to define the task and to label the dataset.

In summary, the confluence of the following trends and opportunities make this an ideal moment to carry out such a project:

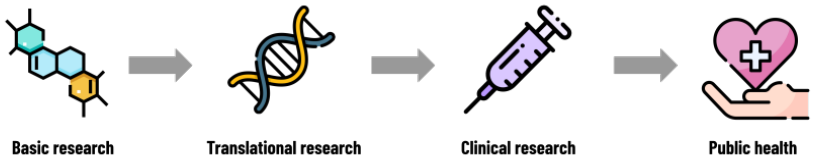
[1] https://ncats.nih.gov/translation/spectrum

- The growing interest in the research phases in the biomedical domain,
- The increasing application of natural language processing techniques in the biomedical domain, and
- The success of pre-trained language models in solving many NLP tasks.

1.2 Problem definition

We have tackled the problem of classifying biomedical scientific literature on the basis of categories in Figure 1.1, defining the borders and formalising the problem, resulting in a framework for defining each phase in detail, even in ambiguous cases. The project defines the BATRACIO task (the acronym of *BAsic-TRAnslational-Clinical research phases classification in bIOmedical publications*), a new BioNLP task. Although different attempts to define the phases in biomedical research can be found in literature (Weber, 2013; Hanney et al., 2015; Flier and Loscalzo, 2017; Fort et al., 2017), there are no universally accepted definitions of these phases. A much more complex challenge arises in assigning individual publications or research projects to each of the research phases.

Figure 1.1: Research phases in biomedical research considered in BATRACIO.



In the initial project proposal, after reviewing the main phases in biomedical research in the literature, we only considered, and defined, the three categories of `basic research`, `translational research`, and `clinical research`, assuming that they were able to cover the whole of biomedical research. Two domain experts attempted to annotate 100 publications in the biomedical domain (randomly extracted from PubMed) with those three proposed categories, redefining the categories iteratively in order to be able to label publications in biomedicine. The experts expressed the need to

add a fourth phase, after clinical research, to cover those scientific publications in biomedical research focusing on topics around health policy, global health, socioeconomic impact on health, population issues, among others. This fourth phase has been named `public health`. These four categories, defined in detail with examples in Chapter 3 and in Appendix C, are presented here:

- **Basic research**, often called *fundamental research*, focuses on scientific exploration and on building new knowledge, and aims to understand fundamental mechanisms of biology, disease and behaviour. For example, in the case of cancer research, basic research asks how or where mutations occur in DNA and how DNA functions in a healthy cell[2].

- **Translational research**, also called *pre-clinical research* (Woolf, 2008) focuses on translating the discoveries from basic research into usability in the clinic, to produce new drugs, devices and treatment options for patients, with a particular focus on applicability. It uses large-scale testing and both animal models and human biological material, such as computer-assisted simulations of drug, device or diagnostic interactions within living systems.

- **Clinical research**[3] seeks to test a specific treatment or procedure, drug, diagnostic or any technology on patients, focusing not only on the biological mechanisms, but also on issues of safety, delivery and protocols for implementation. It includes studies to better understand a disease in humans and relate this knowledge to findings in cell or animal models.

- **Public health** phase involves activities to strengthen public health capacities and services, seeking to provide conditions under which people can stay healthy, improve their health and wellbeing, or prevent the deterioration of their health. For instance, population analyses and retrospective studies, are considered in this phase.

[2] https://blog.dana-farber.org/insight/2017/12/basic-clinical-translational-\research-whats-difference/

[3] https://blog.dana-farber.org/insight/2017/12/basic-clinical-translational-\research-whats-difference/

The objective of *BATRACIO* will be to develop an automatic text classification system able to assign the label corresponding to the futher research phase presented in the article based on the annotated dataset provided by BATRACIO. Providing, for instance, the title and abstract of the following scientific article extracted from PubMed:

Title: *Viral FLIP blocks Caspase-8 driven apoptosis in the gut in vivo.*

Abstract: *A strict cell death control in the intestinal epithelium is indispensable to maintain barrier integrity and homeostasis. In order to achieve a balance between cell proliferation and cell death, a tight regulation of Caspase-8, which is a key player in controlling apoptosis, is required. Caspase-8 activity is regulated by cellular FLIP proteins. These proteins are expressed in different isoforms (cFLIPlong and cFLIPshort) which determine cell death and survival. Interestingly, several viruses encode FLIP proteins, homologous to cFLIPshort, which are described to regulate Caspase-8 and the host cell death machinery. In the current study a mouse model was generated to show the impact of viral FLIP (vFLIP) from Kaposi's Sarcoma-associated Herpesvirus (KSHV)/ Human Herpesvirus-8 (HHV-8) on cell deat h regulation in the gut. Our results demonstrate that expression of vFlip in intestinal epithelial cells suppressed cFlip expression, but protected mice from lethality, tissue damage and excessive apoptotic cell death induced by genetic cFlip deletion. Finally, our model shows that vFlip expression decreases cFlip mediated Caspase-8 activation in intestinal epithelial cells. In conclusion, our data suggests that viral FLIP neutralizes and compensates for cellular FLIP, efficiently counteracting host cell death induction and facilitating further propagation in the host organism.*

The system developed should categorise it as `basic research`, because, according to the annotation guidelines[4], the article aims to understand cell death regulation, in other words, cellular understanding of mechanisms.

[4]Available at Appendix C.

1.3 Proposal and objectives

The main goal of this work is to identify research phases in biomedical research outputs. This project defines a new text classification task in BioNLP. This task consists of automatically classifying scientific literature in the biomedical domain into the 4 phases of the biomedical research value chain. The proposed labels are `basic research`, `translational research`, `clinical research`, and `public health`. It implies a task of text classification, in which each document can be assigned only to a single class. Given a title and description of a scientific publication in the biomedical domain, the system will label the article with the pertinent class.

In order to achieve this, the project aims to create an annotated corpus of biomedical research texts with the categories of the biomedical research phases. It aims to explore how the state-of-the-art BERT-based pre-trained language models perform to solve this task. Different language models pre-trained from the open domain to the specific domain have been explored, and the results have been compared with other machine-learning techniques, to assess whether BERT, and specifically with pre-trained models in the biomedical domain, obtains better results and whether a better generalisation of the context improves results in the task proposed.

To the best of our knowledge, this task has not been addressed automatically before. The definition process has consisted of domain work with three experts to define the annotation guidelines, in which the scope of each category is detailed. These four categories had not been defined in such detail in the literature, since the goal of BATRACIO is to force any article in the biomedical domain to be assigned to one of the categories.

Classifying the scientific outputs of specific funding instruments or scientific publications according to research phases, can help policymakers or funding agencies to better understand what research activities were carried out, mapping stakeholders and their research competencies, and hence to better allocate the resources. This information can also be used to better steer research funds towards proposals that are more appropriate to the ecosystem or would probably have greater impact, avoiding the risk of allowing funding gaps in different research phases and reducing the risk of duplicating efforts. Furthermore, it can also benefit universities in understanding in which areas they are active in practice or to better priority-setting for their research strategy; or for researchers, to most easily extract informa-

tion from large collections of scientific articles.

The general objectives of the project are:

- To define the BATRACIO task and its categories.
- To create an annotated corpus of scientific publications in the biomedical domain for the BATRACIO task.
- To explore whether the problem can be addressed automatically with general-domain systems or if recent pre-trained language models improve on traditional methods.
- To evaluate if domain specific pre-trained language models pre-trained on scientific and biomedical documents outperform general-domain pre-trained language models
- To explore what input data is needed to automatically address the problem
- To explore whether slight modifications can improve the performance of BERT-based systems.

1.4 Document structure

This Master's Thesis is organised as follows:

Chapter 1. Introduction. This chapter presents the main reasons that motivated this work, as well as the problem definition and the current state of the field. Finally, the different contributions of the work are presented.

Chapter 2. State of the art. This chapter describes the discipline in detail, presenting its background up to the present moment. It shows the most frequently used approaches and techniques for solving the most relevant tasks in the domain studied, as well as their main shortcomings.

Chapter 3. BATRACIO, a dataset for classifying texts in biomedical research phases. This chapter describes in depth the method

followed for the creation of the annotated corpus, and presents the main features of the resulting labelled corpus.

Chapter 4. Text classification system based on pre-trained BERT models. This chapter analyses and discusses the proposed text classification systems.

Chapter 5. Evaluation and discussion. This chapter describes the methodology used to evaluate the system and discusses in depth the results obtained in the evaluation of the systems presented in the previous chapter.

Chapter 6. Conclusions and future work. This chapter summarises the different conclusions from this work, and proposes some directions for further work.

Chapter 2

State of the art

This chapter describes in more detail text classification, biomedical natural language processing, and pre-trained language models. It shows the current techniques most commonly used to solve the most relevant tasks of related to the work done in this Master's Thesis, as well as their weaknesses.

2.1 Text Classification

With the explosion of the Internet in the information age, the challenge of classifying massive amounts of data automatically has become fundamental (Li et al., 2020). Automatic text classification is a classic problem in natural language processing, which is defined as the task of labelling natural language texts with a set of predefined tags (Sebastiani, 2001), or assigning classes or categories to text units such as sentences, paragraphs or documents (Minaee et al., 2021). It is also known as *text categorization* (Joachims, 1998) or *topic spotting* (Sebastiani, 2001). Natural text is a rich source of information, but extracting meaningful information from it can be challenging and time-consuming because it is an unstructured data source (Minaee et al., 2021). This task started in the early 60s, but it was in the 90s that it became one of the main fields of interest of the information systems disciplines (Sebastiani, 2001). However, the last decade has seen an even bigger increase in interest in this area by the research community, because of the rise of neural network-based systems (Li et al., 2020). The problem of text classification can be addressed automatically in several ways, from unsupervised or zero-shot learning (Ko and Seo, 2000; Yin et al., 2019a,b), to semi-supervised learning methods (Sun et al., 2020; Li et al., 2021), but in

this project we focus on the study of supervised learning approaches. Text classification is very important in many natural language processing applications (Li et al., 2020), such as: Sentiment Analysis, News Categorization, Topic Analysis, Spam Detection, Question Answering (QA), to mention just some of them. Textual data to be classified can come from multiple sources, such as websites, emails, social media, video transcripts, user reviews, tickets, among others (Minaee et al., 2021).

As described in (Joachims, 1998; Sebastiani, 2001), from a supervised perspective, the problem of classifying text into a predefined set of categories or classes can be formalised as follows. Given a set of categories C, and a collection of documents D which have these categories assigned. The function T maps each document $d_j \in D$ to each class to which the document belongs. For the training data, we know $T(d) \in C$. The information contained in the documents in D may be used by the learning algorithm or system H to train a model able to predict the class to which each new document belongs, given $H(d) \in C$. The mapping offered by $H(d)$ can be used to classify new documents. The main goal is to find the model that maximises the correct predictions.

Text classification can be applied to different levels of granularity, and four levels are pointed out by (Kowsari et al., 2019). At document level, the classifier predicts relevant categories for the whole document. At paragraph level, for a portion of a document. At sentence level, for portions of a paragraph, and at sub-sentence level, for portions of a sentence.

In each text classification problem, different constraints guide how and how many categories or classes can be assigned to each document, and this is a fundamental issue in the problem conceptualisation. Sebastiani (2001) proposes to differentiate between *single-label* and *multi-label* classification. *Single-label categories*, also referred to as non-overlapping categories, are those problems in which each document can only be labelled with one category. The most basic problem is a binary classification problem, in which the documents are labelled with a boolean category. This is the case of a standard spam detection system, which decides if the new document is or is not spam, or a system that classifies social media posts according to whether they are sexist or not. In these two examples, there are only two categories and only one can be assigned to each document. The other case of single-label is when there are more than two categories, but each document

only belongs to one category. This is sometimes called *multi-class* text classification. An example of multi-class is the classification of news with their main topic, within a possible set of topics such as sports, finance, politics, international and society. By contrast, if the document can be labelled with more than one class, this is a *multi-label* problem, also known as overlapping categories.

2.1.1 Evolution of text classification techniques

Text classification techniques have evolved a lot in the last twenty years. The first techniques were based on rules, then on traditional machine learning algorithms, and, in the last decade, techniques based on deep learning have been particularly outstanding, because of their successful results (Li et al., 2020). Automatic text classification approaches can be divided in two major paradigms (Minaee et al., 2021):

- *Rule-based* methods.
- Machine learning (data-driven) based methods, where we could differentiate between:
  - Traditional methods, or *Shallow learning* (Li et al., 2020).
  - Deep learning, or neural approaches (Minaee et al., 2021).

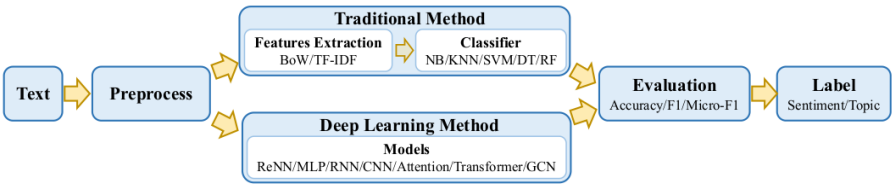
Rule-based methods, or *knowledge-based* methods (Cohen, 2014), consist of differentiating categories on the basis of a set of pre-defined rules. Those rules can be defined by hand, although some automatic technique have been also proposed to propose rules from training data (Aubaid and Mishra, 2020). Generally, these rules follow the form if "condition" then "category", based on symbolic representations of knowledge. These systems are easily interpretable. However, the task of defining those rules requires deep domain knowledge (Minaee et al., 2021) and intensive manual work. Therefore, these systems do not obtain the best results, because of their lesser learning capacity, and complex domains can be specially challenging for them (Thanaki, 2017) because they cannot capture complex messages, inexplicit references to the categories and hidden patterns related to the domains.

In contrast, machine learning approaches learn from the observation of data, using pre-labeled data to learn inherent relations between texts and their labels to classify (Minaee et al., 2021). But unlike images and numerical

data, text data require NLP techniques to process the texts well (Li et al., 2022), such as text segmentation and tokenization, lemmatization and word sense disambiguation.

Most *traditional* machine-learning methods follow a two-step approach, based on, feature extraction, which is fundamental for the effectiveness of the method, and then to feed a classifier with those features. This requires feature analysis to obtain good performance (Minaee et al., 2021), but how to reduce the features efficiently is challenging, because the number of features can increase the computation cost heavily. Bag of Words (BoW) (Zhang et al., 2010), N-gram (Cavnar and Trenkle, 2001), or Term Frequency Inverse Document Frequency (TF-IDF) (Sebastiani, 2001), among their variants, are widely used handcrafted extraction methods which allow us to represent words in a mapping array of tokens by relevance and frequency. Traditional methods include statistic-based classifiers, such as Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), or Random Forests (Minaee et al., 2021; Li et al., 2022; Kowsari et al., 2019). These methods are much more accurate and robust than earlier ones based on rules. However, they rely on feature extraction, which is costly and time-consuming (Li et al., 2022). Furthermore, this can limit the portability or generalisation of the systems for further applications to new datasets or domains (Wei et al., 2020).

Figure 2.1: Flowchart of text classification based on machine learning methods, extracted from (Li et al., 2022).



In the last years, deep neural network-based techniques have stood out because of their simplicity, reducing the costs of manual feature extracting, higher processing efficiency, and, in general, because they have managed to match or improve state-of-the-art results in many NLP tasks (Wu et al., 2019). Deep learning methods include feature extraction in the model fitting process by learning a set of nonlinear transformations that allow the mapping of features directly to outputs (Li et al., 2022). One of the main advantages

of these techniques is that they can be trained on unstructured data and can learn feature representations directly from the input text without much manual intervention or prior knowledge (Qiu et al., 2020).

Some of the most widely used algorithms in NLP have been MultiLayer Perceptron (MLP), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Graph-based Neural Networks (GNNs), attention mechanism, and Pre-trained Language Models based on Transformer architecture (Qiu et al., 2020; Li et al., 2022). RNNs, and their improvements such as LSTM or biLSTM, can capture distant dependencies in sequence by recurrent computations. However, they compute sequentially and cannot be parallelised, and this makes it challenging to build models with more layers and parameters efficiently, because they take long time to train and take a lot of memory. On the other hand, CNNs can automatically extract features from texts, applying convolution filters of different sizes. They capture local information well, but presents more weaknesses capturing long-distance information. GNNs can better capture syntactic structural information in a text by constructing a graph from it. Attention mechanism allows the model to pay different attention to specific words or sentences, identifying the importance of each word or sequence for the classification. Attention mechanism improves performances and allows better interpretability than CNNs and RNNs. Nevertheless, Transformer seems to capture better long dependencies in text, improve in computationally and take less time to train. It treats text as a fully-connected graph with attention features between different words by self-attention, which can extract features and relations between words efficiently based on self-attention, which solves short-term memory problems (Li et al., 2022).

Pre-trained language models based on Transformer (Vaswani et al., 2017) learn global semantic representation from a large dataset and successfully solve many NLP tasks. They generally use unsupervised methods on large datasets, and then they are adapted to a downstream task, without having to train the whole model from scratch. Pre-training allows the model to learn to understand semantics and context. Some of these are ELMo (Peters et al., 2018), GPT (Radford et al., 2018) and BERT (Devlin et al., 2019). BERT improved performance on eleven NLP tasks (Devlin et al., 2019). It applies bidirectional encoding, and pre-training strategy and after fine-tuning to downstream text classification task by adding a linear classifier in the output

layer. It is the first fine-tuning-based representation model that achieves state-of-the-art results for several NLP tasks. It shows how adapting a pre-trained language model to a downstream task can be really powerful (Sun et al., 2019), because initialisation with the general representation allows starting with a general representation and adapting it to the task, instead of initialising with random values (Qiu et al., 2020).

Although methods based on deep learning are particularly powerful, they also have several limitations. For instance, they require large amounts of data to achieve high performance and large computational resources are needed. Other major challenges when it comes to deep learning methods are data explainability and interpretability. However, models based on self-attention can provide a bit more of these aspects based on the relationship between words. Although, traditional methods are much clearer when we want to understand why and how they work well, despite being less robust and effective (Li et al., 2022).

2.2 Text classification in BioNLP

2.2.1 Introduction to BioNLP

For more than twenty years, the number of published biomedical literature, clinical reports, clinical trials and electronic health records, has experienced a continued growth (Naseem et al., 2021; Huang and Lu, 2015; Else, 2020; Cohen, 2014). In fact, at present, more than 3,000 publications in the biomedical domain appear every day (Kim et al., 2018).

Identifying relevant information in biomedical literature to review research about a specific protein or exploring new discoveries in a specific discipline, is fundamental for any researcher in health and life sciences. In the same way, health practitioners need to review available electronic health records to carry out their medical practice (Huang and Lu, 2015).

In view of this situation, one could think that the growing amount of available information and evidence is very useful to foster new discoveries in biomedicine and to achieve improvements in disease diagnosis. However, accessing and reading this large and growing amount of documents manually would be extremely time-consuming, and therefore, not feasible for humans. So, we face a contradiction, we have more and more documents than ever, but the access to them is increasingly harder.

Natural language processing (NLP) aims to study human language so that computers can understand natural language as humans do, and its applications range from identifying parts of speech of words to answering questions. Unfortunately, NLP techniques, or systems, that work in general domain texts, do not work as well in specific domains. The application of NLP in the biomedical and clinical domains is known as biomedical natural language processing (BioNLP), and in the last twenty years, the work produced in this field has been growing considerably (Friedman et al., 2002; Chapman and Cohen, 2009; Huang and Lu, 2015; Wu et al., 2019; Percha, 2021).

The applications of BioNLP are able to contribute to different dimensions of healthcare and biomedical research. Some impact areas of this field are contributing to the diagnosis with clinical decision support systems, helping patients to understand their own medical records, and for public health and biomedical research improving the access to the biomedical literature (Leaman et al., 2015).

Clinical and biomedical language have particular features which make their automatic processing more challenging. As languages in specialized domains or *sub-languages* (Harris, 1991), they have their own grammatical features and terminology (Friedman et al., 2002). Cohen (2014) points out some potential challenges of processing language in this domain: the negation detection which is basic to understand the meaning of a text, polysemic words, the lack of grammatical consistency and punctuation of clinical notes, or the frequent use of parenthesis and abbreviations in biomedical literature which introduce additional ambiguity. However, the major difficulty of this field could be the *knowledge* of the world, in other words, the information behind the words which is expected to be known or understood from the context. Omitted information is especially problematic for BioNLP because a system must have additional knowledge to be able to gather all the implicit information (Friedman et al., 2002). In fact, many of these documents are not clearly accessible or understandable by humans without adequate domain expertise/knowledge.

The major types of documents this research field has studied are mainly clinical documents and biomedical literature (Cohen, 2014). Nevertheless, other sources such as social media or web content are also used. An application of BioNLP on social media could be pharmacovigilance (Harpaz et al.,

2014); an example of this would be the monitoring of experiential medication effects reported by consumers on Twitter (Zhu and Jiang, 2021). And an example of web content could be the exploration of automatic question-answering systems based on health questions from FAQ sections of NIH[1] websites (Abacha and Demner-Fushman, 2016).

Chapman and Cohen (2009) points out as main reasons behind the development of this field, together with the exponential year-on-year growth of the number of scientific publications, the free access to databases such as PubMed/MEDLINE (Canese and Weis, 2013) or PubMedCentral and to a wide variety of corpora, semantic resources and ontologies, such as UMLS (Bodenreider, 2004) or Gene Ontology (Ashburner et al., 2000). Furthermore, in recent years, the advancements of deep learning in natural language processing have fuelled the development in a range of BioNLP tasks (Demner-Fushman et al., 2021).

The interest in deep learning techniques due to their good performance is also observable reviewing research contributions of the last few years in the *Biomedical Natural Language Processing Workshop (BioNLP)*(Demner-Fushman et al., 2018). Transformer-based pretrained language models are explored in the overwhelming majority of the papers in the *Proceedings of the 20th Workshop on Biomedical Language Processing* to solve all kinds of fundamental NLP tasks in the biomedical domain (Demner-Fushman et al., 2021). Deep learning approaches usually require a lot of annotated data, and there are limited labelled data because of the high annotation costs. Moreover, the addition of external knowledge of the domain to the models is still a major challenge (Demner-Fushman et al., 2021).

Evaluation conferences have played a fundamental role by proposing new specific shared tasks and providing frameworks for the evaluation of the systems and techniques. There has been a growth in the number of tasks related to BioNLP or scientific documents. Perhaps the fundamental tasks that have been most present in shared tasks are those related to entity identification, but there has been a great variety of others. Related to entity recognition and normalisation, SemEval proposed the *Analysis of Clinical Text Task* (Elhadad et al., 2015) in order to map entities and acronyms to UMLS concepts in clinical texts, or, for instance, the *Information Extraction from Noisy Text Task* (Goeuriot et al., 2021a) in CLEF eHealth.

[1] National Institute of Health of the United States.

Also a NER in Spanish, there was the *PharmaCoNER task* (Gonzalez-Agirre et al., 2019) in BioNLP, with the goal of identifying pharmacological substances, compounds, and proteins. IberLEF has also fostered NER-related task such as *MEDDOPROF Task* (Lima-López et al., 2021) which aimed to detect professions and occupations from medical texts, or the *Disability annotation on documents from the biomedical domain (DIANN) task* (Fabregat et al., 2020) that included NER and negation detection. Other shared tasks also betting on RE such as *Extraction of Drug-Drug Interactions from Biomedical Texts Task* (Segura-Bedmar et al., 2013) in SemEval, or the multilingual *eHealth-KD 2021* (Guan and Liu, 2021) in IberLEF. In relation to Information Retrieval, there have been the *RDoc Task* (Anani et al., 2019) in BioNLP for exploring IR systems in the field of neuroscience, or the *Consumer Health Search* (Goeuriot et al., 2021b) in CLEF eHealth. Another task in BioNLP, *MEDIQA 2021 Task* (Ben Abacha et al., 2021) aimed for different summarization tasks of medical texts. QA has been studied in both *Biomedical Semantic Question Answering Task* (Telukuntla et al., 2020) and *Synergy Task* (Nentidis et al., 2021), focused on answering questions for COVID-19, in CLEF BioASQ. Some others more oriented to scientific literature in SemEval could also be mentioned, such as the *Statement Verification and Evidence Finding with Tables Task* (Wang et al., 2021) or the *Extracting Keyphrases and Relations from Scientific Publications Task* (Augenstein et al., 2017).

On the other hand, benchmark datasets collect reference tasks to evaluate models or systems in different basic BioNLP tasks and are quite significant in the evaluation of neural language models, to assess their capacity to transfer learning from unlabelled corpora to specific BioNLP downstream tasks. Two of them are the BLUE and the BLURB benchmarks. The BLUE (Biomedical Language Understanding Evaluation)[2] benchmark is proposed by the NCBI, and it consists of five different tasks based on pre-existing datasets from shared tasks. The BLURB (Biomedical Language Understanding and Reasoning Benchmark)[3] is inspired by previous resources, and contains thirteen evaluation datasets for six diverse tasks.

[2] $https://github.com/ncbi-nlp/BLUE_Benchmark$
[3] $https://microsoft.github.io/BLURB/$

2.2.2 Text classification methods in biomedical domain

Text classification on biomedical and medical records is specially challenging compared to general domain, owing to imbalanced dataset, misspelling, acronyms and abbreviations, negations and semantic ambiguity (Mascio et al., 2020; Cohen, 2014; Mujtaba et al., 2019). Although deep neural network-based approaches have allowed relevant results to be obtained, the creation of generalizable classification models able to obtain good results in both clinical and medical texts is still a complex task, because it is difficult to cover all existing concepts (Qing et al., 2019; Patel et al., 2018). Pre-trained models of language have been particularly salient in recent years, but they also present big limitations in the biomedical domain, and a few initiatives have explored how to enrich pre-trained language models with biomedical and clinical knowledge (Gu et al., 2022; Wada et al., 2021; Huang et al., 2020).

In BioNLP several techniques have been used for text classification, ranging from rule-based approaches, machine learning and deep learning, including pretrained language models. Rule-based approaches have been widely used, and they are currently used in research and domain applications, given their capacity to incorporate domain knowledge (Mujtaba et al., 2019), and because of the challenge of extracting meaningful features. Although their performance can be lower than other methods, rule-based approaches allow a much easier involvement of experts. However, rules can become very complex.

However, as Mascio et al. (2020) points out, in the medical domain, rule-based systems continue to predominate. Moreover, there is no consensus on which word representation approach is the most adequate for specific classification tasks in the domain (Mascio et al., 2020). In the medical domain, machine learning-based techniques tend to improve the results of rule-based systems, but machine learning-based systems in most cases require a large annotated corpus (Patel et al., 2018), and especially in these domain datasets which are small and limited.

Traditional machine-learning methods have been used to learn from features extracted from documents, but the number of features can grow fast and suffer from a problem of data sparsity. Selected features cannot cover all linguistic variants. This problem can be even be more prominent in a domain with small datasets. To go beyond this issue, more complex fea-

tures are used when dealing with biomedical or clinical documents, such as citation information (Yepes et al., 2015) or information from ontologies (Sanchez-Pi et al., 2016).

Deep learning-based methods, like CNNs (Yao et al., 2018b) or RNNs (Burns et al., 2019), have improved traditional methods in biomedical and clinical domains. Especially in combination with word embeddings, including Doc2Vec (Koutsomitropoulos and Andriopoulos, 2020), Word2Vec (Prabhakar and Won, 2021) or FastText (Agibetov et al., 2018), they have been valuable. However, they are still limited because they are generally trained on general-domain documents, but proposing word embeddings pre-trained in biomedical documents to mitigate this problem have also been explored (Agibetov et al., 2018; Yijia et al., 2019).

Nevertheless, hybrid approaches combining different methods seem to be a prominent trend in the domain, because most new deep learning-based methods still lack knowledge of the domain. Among these, we can find proposals combining convolutional neural networks with rule-based features (Yao et al., 2018b), hybrid biLSTM (Shen and Zhang, 2020), or incorporating knowledge from the UMLS ontology while training a pre-trained language model (Yuan et al., 2021).

Document classification tasks are frequent in the literature, although they are less present than other tasks, such as entity recognition or relation extraction. Following these, we present some diverse examples of document classification in biomedical and clinical domain proposed in recent years.

- **Hallmarks of Cancer (HOC)** (Baker et al., 2016), this task consists of classifying scientific publications according to the Hallmarks of Cancer taxonomy. A dataset with 1852 PubMed publication abstracts manually annotated by experts according to the Hallmarks of Cancer taxonomy is provided. The taxonomy consists of 37 classes in a hierarchy. Zero or more class labels are assigned to each sentence in the corpus. This is part of BLURB and BLUE benchmarks.

- **MeSH Class Labels** (Cohan et al., 2020) formed by 23k documents, which consists of classifying medical scientific documents according to 11 top-level disease classes extracted/derived from the Medical Subject Heading (MeSH) vocabulary, where each document has to be assigned to a category.

- **Automatic ICD[4] Coding** (Pascual et al., 2021), they use medical notes from the MIMIC-III dataset (Johnson et al., 2016), which consists of medical notes labelled with ICD codes from the ICD-9 taxonomy, a multi-label classification, where each document as an average of 13.15 ICD Codes.

- **Claim Detection in Biomedical Tweets** (Wührl and Klinger, 2021), offers a biomedical claim detection dataset in social media for predicting if a tweet contains a claim and, later, to distinguish between explicit claim, implicit claim, and non-claim.

2.3 Pretrained Language Models

2.3.1 From static to dynamic word representation

Word and word sequence representations have been a fundamental focus of research in NLP (Wang et al., 2020). Some of the most widely-used text representation methods are TF-IDF or one-hot vectors. These are high-dimensional, and suffer from data sparsity, and they cannot encode syntactic and semantic information (Kalyan et al., 2022). Words are represented at a surface level, and a good pre-processing pipeline has to be implemented to consider derived words as the same. Hence, the need to generate low-dimensional embeddings that can encode information at the semantic and syntactic level. In other words, to generate embeddings able to encode word or sequence meaning in a vector (Qiu et al., 2020).

Some of the main problems were how to represent sequences considering the context of the words, learning universal language representations and not training the language representation from scratch on the datasets, which can reduce the generalisability of the system drastically, especially with small datasets, for instance, caused by the out-of-vocabulary problem. Most datasets for NLP tasks are rather small, except for machine translation. This is one of the main limitations of using neural network-based approaches, because the large number of parameters in deep neural networks can overfit on small datasets. It can also reduce the portability and generalizability of systems (Qiu et al., 2020).

However, pre-trained language models (PLMs) aim to learn universal

[4]International Classification of Diseases

language representation from large corpora, which can then be applied to a downstream task, thus avoiding the need of training models for a specific task from scratch. These approaches have had a great impact in the last decade in NLP (Qiu et al., 2020). We can differentiate between *contextual* and *non-contextual* embeddings[5].

Pre-trained language models of first generation or *non-contextual*, such as Word2Vec or GloVe, are able to get embeddings at word level, capturing semantic meaning of words, but their word representation does not capture information about context of words. It makes difficult, for example, polysemic disambiguation, identification of syntactic structures or of semantic roles in a text (Qiu et al., 2020). One of the main limitations of this kind of *non-contextual* embeddings is that they are static, meaning that the representation of a word does not change according to the context, and therefore they fail with polysemic words. Another common problem is that they mostly suffer from out-of-vocabulary words (Qiu et al., 2020). One of the first attempt to obtain generic word embeddings trained with unlabelled data that could be useful for other tasks was (Collobert et al., 2011). Then, (Mikolov et al., 2013) proposes Word2Vec, which implements the architecture of Continuous Bag-of-Words (CBOW) and Skip-Gram, two shallow architectures based on two-layer neural network, demonstrating that deep neural network were not necessary to obtain good embeddings. This implementation is one of the most popular PLMs for NLP tasks (Qiu et al., 2020). Also there is GloVe (Pennington et al., 2014), another popular implementation of word embeddings that computes word to word co-occurrence on a large corpus to obtain global embeddings, and FastText (Bojanowski et al., 2016), which proposed to encode sub-word information to face the recurrent out-of-vocabulary problem. Although using these PLMs the representation of words is already trained, being context independent, the remaining part of the model has to be trained from scratch. Furthermore, most NLP tasks go beyond the level of word, dealing with problems at sentence, paragraph or document level. Other contributions, such as doc2vec (Le and Mikolov, 2014) or text2vec, attempt to address learning embeddings at sentence or document level at a fixed dimensional rather than using word-level representation. However, they are not more effective than word-level embeddings

[5]In (Qiu et al., 2020), *non-contextual* are named *first generation* embeddings; and *contextual*, *second generation* embeddings.

(Qiu et al., 2020).

Second generation PLMs or *contextual* embeddings learn the context of words and go generally beyond word level; some examples are ULMFiT, CoVe, ELMo, OpenAI GPT or BERT[6]. Contextual embeddings address the context-dependent nature of words and issues such as polysemy, different meaning of words in different contexts. More modern versions of PLMs are trained on larger scale corpora, and on more powerful architectures, such as Transformer, and based on new pre-training tasks. The fine-tuning technique to adapt PLMs to a downstream task proposed by ULMFiT and BERT is the mainstream approach (Qiu et al., 2020).

Deep neural networks have made it possible to greatly increase the number of parameters. Likewise, to prevent overfitting, and to be able to make use of the potential of deep learning, a much larger dataset is needed. Unfortunately, building sufficiently large annotated datasets is very expensive. On the other hand, unlabelled corpora are much easier to obtain. Thus, in order to take advantage of all these data, learning a good representation of these corpora and then applying this representation to a downstream task allows to obtain significant performance improvements. Pre-training allows to get universal representations of the language beyond the limited training set, it allows a good initialisation of the model which in turn provides a higher generalisability which allows the acceleration of the matching with the task, and to avoid overfitting in small datasets, what can be seen as a kind of regularisation. (Qiu et al., 2020).

2.3.2 Bidirectional Encoder Representations from Transformers (BERT)

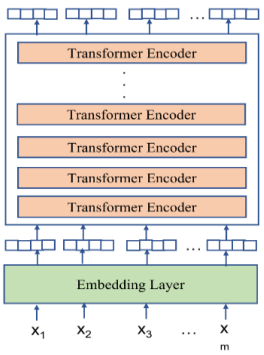
From the combination of *transfer learning*, as proposes ULMFiT (Howard and Ruder, 2018), which shows how training LSTMs on large corpus could build state-of-the-art text classifiers, with the Transformer encoder architecture (Vaswani et al., 2017), appears BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). Combining these two elements, these models do not have to train task-specific architectures from

[6]So far, many new proposals have appeared, such as RoBERTa, BigBird, AlBERTa, or XLM. They are not considered in this work, since although they improve the state of the art by increasing the number of parameters, size of pre-training data, pre-training tasks, among others. They do not change the encoding paradigm, and there are not many models available in the biomedical domain and scientific literature, to explore in this project.

scratch (Tunstall et al., 2022). BERT is a state-of-the-art language representation model proposed by (Devlin et al., 2019), and it supposes a major advance in pretrained language models in NLP, as it is the first one to propose pretrained deep bidirectional representation jointly conditioning on both left and right contexts in all layers (Devlin et al., 2019). One of the main contributions of BERT is that demonstrates the relevance of bidirectional pre-training for language representation. So far, systems could not process texts by incorporating contextual information from both sides of the text. It attempts to solve the limitation of unidirectional language models or pseudo-bidirectional as their predecessors ELMo and GPT did not propose a bidirectional architecture, a token only serves previous tokens with self-attention layers of Transformers (Vaswani et al., 2017), since to generate good representations requires context on both sides, by using Masked Language Modelling pre-training task. This works by randomly masking some tokens from the inputs, and the goal is to predict the original token by the context. Additionally, BERT is also trained on Next Sentence Prediction, to predict the next sentence pre-training pairs of sentences.

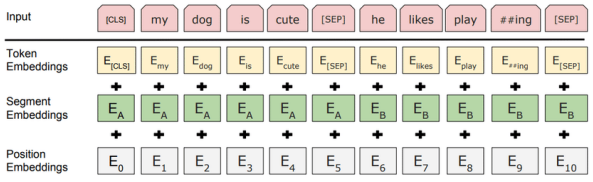
BERT is composed by a multi-layer bidirectional Transformer based on the original implementation by (Vaswani et al., 2017), and it is pre-trained on a large corpus for masked word prediction and next sentence prediction tasks. BERT proposes a fine-tuning approach to adapt the PLMs to a downstream task. This is the first fine-tuning based representation model achieving state-of-the-art results on several NLP tasks, improving task-specific architectures. They demonstrate that self-supervised pretraining on large corpora can reduce heavily-engineered task-specific architectures.

Figure 2.2: Architecture of PLMs like BERT, diagram extracted from (Kalyan et al., 2022).



Vaswani et al. (2017) proposes the Transformer architecture as an encoder-decoder framework. This architecture outperforms RNNs and CNNs in machine translation, in quality and in computational cost (Tunstall et al., 2022), because it can be parallelized. However, PLMs like BERT are based on a combination of embeddings and transformer encoder layers (Kalyan et al., 2022). The embedding layer has different sub-layers, each of them encodes a specific type of information. In the case of BERT, the embedding has three layers, and provides information about token, segment and position, for each input token. The input vector is computed by summing the different embeddings. Input tokens are generally introduced in a sub-word level, because it can overpass the out-of-vocaublary problem and to capture additional information about words. For instance, BERT implements WordPiece (Wu et al., 2016), with 30,000 token vocabulary. General-domain BERT is pretrained on BooksCorpus and English Wikipedia.

Figure 2.3: BERT embedding layer, extracted from (Devlin et al., 2019).

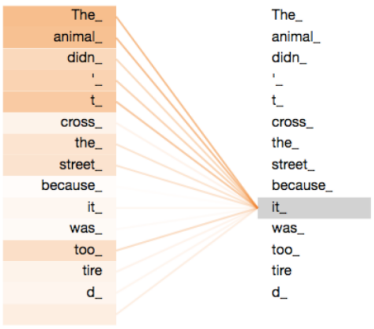


The transformer encoder layers encode contextual information of each input token by applying self-attention mechanism. A sequence of transformer encoder layers are applied to encode complex language information from the input token. The transformer encoder layer consists of a combination of Multi-Head Self Attention, Position-wise Feed Forward Network, Add and Norm (Devlin et al., 2019; Kalyan et al., 2022).

A fully-connected self-attention model can capture the relationship between words and let the model learn the structure by itself, computing the connection weights between words dynamically with the self-attention mechanism (Qiu et al., 2020). It is a better alternative than convolutional and recurrent layers to encode global contextual information (Kalyan et al., 2022). To avoid that each token attends to itself, multiple self-attention layers are applied, each with different weight matrices to attend to multiple positions while encoding a word (Kalyan et al., 2022).

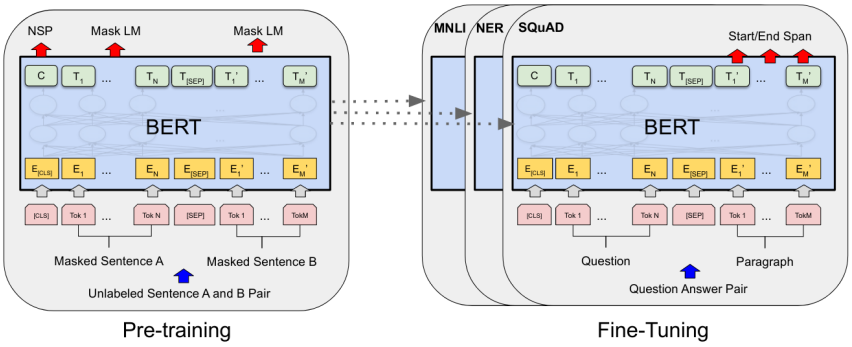
The proposed workflow in the original paper for BERT has to two steps,

Figure 2.4: Self-attention example, extracted from *Illustrated transformer*, http://jalammar.github.io/illustrated-transformer/.



pre-training and fine-tuning (Devlin et al., 2019; Rogers et al., 2020). The advantage of fine-tuning is that only few parameters need to be learned for the downstream task. During fine-tuning the parameters are initialised with pre-trained values. It has a unified architecture, only few differences between pre-trained architecture and final downstream architecture (Devlin et al., 2019). Different studies have explored why BERT works well and its performance, and (Rogers et al., 2020) give an overview of them.

Figure 2.5: Overall pre-training and fine-tuning procedures for BERT from (Devlin et al., 2019).



2.3.3 Adaptation of PLMs to Text Classification Tasks

There are two main, popular approaches to train the model for text classification: the feature-based approach (Devlin et al., 2019; Tunstall et al., 2022) and the fine-tuning (Devlin et al., 2019; Sun et al., 2020). In the feature-based approach, model parameters are frozen, and it would consist

of using the hidden states as a low-dimensional representation of texts and training a classifier on top (Tunstall et al., 2022). Fine-tuning consists of training the model end-to-end, updating the parameters of the pretrained model.

Simple fine-tuning for text classification consists of stacking a linear classifier on top of the last hidden state of the pre-trained BERT. A simple softmax classifier is added at the end of the model to predict the probability of a label. All the parameters from the model are fine-tuned by maximizing the log-probability of the correct label (Sun et al., 2019).

The feature-based approach could train efficient and fast, because parameters don't have to be updated. In fact, this approach could work specially good with small datasets, given that there are not many samples to train, and the pre-trained knowledge can better generalize the task, and the small training could quickly overfit or produce catastrophic forgetting. For instance, the text classification on the Specter model (Cohan et al., 2020) which is pre-trained on scientific literature by using citation information to reduce loss, outperforms the fine-tuned version of SciBERT (Beltagy et al., 2019) in topic classification in scientific and biomedical domain, only by training a linear SVM as hidden state as only feature. However, Devlin et al. (2019) find fine-tuning adaptation performs better than feature-based adaptation.

In (Sun et al., 2019), they investigate how to efficiently use BERT for text classification tasks, exploring three different ways of fine-tuning to improve performance in text classification. Firstly, during the fine-tuning to the target task, they explore the possible contribution of each layer in the capture of different levels of semantic and syntactic information; given that lower layers encode more general information, they fine-tune with different learning rates in each layer. Secondly, they explore if further pre-training of the general domain model to data in the target domain could improve. Finally, they explore if multi-task fine-tuning can benefit single-task fine-tuning. To avoid the overfitting problem, they observe the need of an optimizer with an appropriated learning rate. They use the uncased BERT-base model. They fix a batch size of 24 during fine-tuning, and default parameters. They empirically fix the max number of epochs to 4, and save the best model on the validation set for testing. Dealing with long texts, because BERT cannot process texts with a length larger than 512, they explore dif-

ferent ways of dealing with long texts, and see that head+tail truncation achieves the best performance. They also corroborate that features from the last layer of BERT give the best performance. However, they observe that *catastrophic forgetting* (McCloskey and Cohen, 1989) is an important problem in transfer learning, which consists of erasing the pre-trained knowledge during the learning of the new knowledge. They observe that lower learning rates, as 2e-5, are necessary to overcome catastrophic forgetting (Sun et al., 2019). Further, pre-training and in-domain further pretraining, are useful to improve performance for the text classification task, but cross-domain pre-training does not benefit. They see that BERT can also improve the task with small-size data (Sun et al., 2019).

Arslan et al. (2021) explores multi-class text classification, they point out that text classification tasks have been poorly studied in the literature by evaluating pre-trained language models. They use different pre-trained models to classify documents in the financial domain. Even with vocabulary adaptation, they conclude that no substantial improvement is achieved in this domain by using in-domain models compared to generic models.

2.3.4 BERT models in BioNLP

BioNLP is the field of natural language processing which tackles the particularities of clinical and biomedical documents. Natural language processing in this domain is especially challenging due to its complexity, specifically because of many in-domain specific words (Cohen, 2014). Performance of general-domain PLMs in BioNLP tasks has been limited in some cases, researchers focused on building PLMs to face biomedical texts(Kalyan et al., 2022), and pretraining PLMs on biomedical corpora improves performance (Lee et al., 2019; Gu et al., 2022). Transformer-based pretrained language models are explored in the overwhelming majority of the papers in the *Proceedings of the 20th Workshop on Biomedical Language Processing* to solve all kinds of fundamental NLP tasks in the biomedical domain (Demner-Fushman et al., 2021).

Some BERT models developed to tackle texts in biomedical domain following different domain-adaptation strategies and datasets, such as BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), BioMedBERT (Chakraborty et al., 2020), OuBioBERT (Wada et al., 2021), PubMedBERT (Gu et al., 2022) or BlueBERT (Peng et al., 2019). However, each one implements a

different strategy for specific-domain adaptation or pre-training, and each has shown different success. The most popular sources for pre-training on large corpora are PubMed and PubMedCentral, which allow open access to biomedical literature (Kalyan et al., 2022). However, general-domain BERT should perform better in the biomedical domain than other general-domain architectures, because it has been pre-trained on the medical subset of Wikipedia (Sushil et al., 2021).

BioBERT (Lee et al., 2019), SciBERT(Beltagy et al., 2019) and BlueBERT (Peng et al., 2019), are further pre-trained from the general domain BERT, initializing from its weights. BioBERT (Lee et al., 2019) is initialised from weights of general-domain BERT (pre-trained on English Wikipedia and BooksCorpus), and pretrained on PubMed abstracts and PMC full-text articles. They keep the same vocabulary of BERT pre-trained on general domain, BaseVocab, and any new in-domain word is represented in the model. BlueBERT (Peng et al., 2019) is pre-trained on PubMed abstracts and clinical notes, and demonstrates the importance of pre-training among different text genres. It improves BioBERT in all BLUE benchmark task test sets (Peng et al., 2019). SciBERT (Beltagy et al., 2019) also starts from general-domain weights, but includes a new vocabulary in scientific domain, SciVocab, which only overlaps 42% with BaseVocab. This model is adapted to the domain on a random sample of 1.14M full text papers from Semantic Scholar, consisting of 18% in computer science and 82% in the biomedical domain. PubMedBERT (Gu et al., 2022) is one of the most successful and widely used in BioNLP tasks, and has been pre-trained from scratch on PubMed abstracts, and it demonstrates that domain-specific pre-training outperforms mixed pre-training as BioBERT does. OuBioBERT (Wada et al., 2021) applies simultaneous pre-training on a small dataset in the medical field with general-domain documents, based on a well-balanced pre-training by up-sampling instances derived from a corpus appropriate for the target task, and also with an amplified vocabulary. This model improves document classification tasks in the medical domain and provides successful results in English and other languages.

Including additional information during the pre-training seems to be a good strategy for improving performance of PLMs. For instance, LinkBERT (Yasunaga et al., 2022) and Specter (Cohan et al., 2020) include citation information as additional input. LinkBERT (Yasunaga et al., 2022) in-

cludes document links to improve adding salient background knowledge to the language model, these links can bring knowledge that is not in a single document. The BioLinkBERT (Yasunaga et al., 2022) offers a new state-of-the-art result on some BioNLP tasks. This adds the Document Relation Prediction as a pre-training task. By using the same dataset as in PubMedBERT it improves in performance, adding the citations between articles as document links. Specter (Cohan et al., 2020) is pre-trained on scientific literature, Semantic Scholar corpus (Bhagavatula et al., 2018), and initialised from SciBERT (Beltagy et al., 2019) weights. This is proposed for document-level representation tasks, simply by a feature-based approach instead of task-specific fine-tuning.

Adding additional knowledge to these models is crucial to solve BioNLP tasks successfully, and it has been effective in other domains. KeBioLM (Yuan et al., 2021) explores incorporating knowledge into language models in biomedical domain by adding knowledge from the UMLS knowledge base, and adds the two new pretraining tasks, namely entity linking and entity detection. In their experiments, the authors demonstrate successful results in entity recognition tasks in the biomedical domain. However, how to add or improve knowledge in pre-trained language models is challenging. Sushil et al. (2021) explore different ways of enriching clinical domain knowledge in BERT models, but they do not produce significative improvements with respect to the baseline BERT models. DiseaseBERT (He et al., 2020) integrates knowledge about diseases such as treatment, diagnoses, among other aspects to the BERT model. They do this by adding the new pretraining task of Disease Knowledge Infusion training.

Some other models have been pre-trained on clinical narratives, such as ClinicalBERT (Alsentzer et al., 2019) which is initialised from BERT-base and BioBERT, although this does not achieve state-of-the-art results in clinical tasks. BERT-EHR (Li et al., 2019b) is also pre-trained on electronic health records, and it includes code, procedure, medication, age, and gender in the embedding layer. Some others are pre-trained for dealing with other sources type in biomedical domain such as BERTTweetCovid19 (Müller et al., 2020), or BioRedditBERT (Basaldella et al., 2020), for being applied to text from social media.

Although text classification is not one of the most popular tasks in BioNLP, some tasks and articles can be found that explore the use of PLMs

in the domain. In (Pascual et al., 2021) explore automatic International Classification of Disease (ICD) coding, which is a multi-label classification problem. However, they have to deal with the length limitation of these BERT-based models, which they can only process a maximum length of 512 tokens, and clinical notes usually exceed this maximum input length. Given that these limitations do not allow the fine-tuning, they go for feature extraction and classifier, using PubMedBERT. They pre-process texts, converting all text to lower case and removing all numbers, but they do not remove infrequent words, because BERT does not suffer from out-of vocabulary terms. They also apply early stopping on the validation set for avoiding overfitting. In fine-tuning, they observe that 6 epochs work better than only 3. In the case of discharge summary, front truncation yields the best performance, although mixed also have a competitive score. In contrast to most NLP tasks, Transformer architecture is not the state-of-the-art system in assigning ICD codes.

In (Wührl and Klinger, 2021), they aim to text classification of claims in biomedical tweets. They experiment with combinations of the recommended fine-tuning hyperparameters from (Devlin et al., 2019), which consist of batch size, learning rate, and number of epochs, and they use those with the best performance on the validation data. Further, they oversample the minority class of implicit claims to achieve a balanced training set. In order to guaranty enough comparability, they oversample. They use default parameter values. In their experiments, the more complex models, such as BERT or LSTM, do not outperform the linear models. They attribute this to the fact that complex models are not able to learn from the training set and to the small size of the dataset.

Reviewing some articles on similar tasks in biomedical domain, some of the main challenges can be identified. For instance, in relation extraction tasks, different variants of BERT may achieve the best results, and fine-tuning the models is preferable over freezing the layers of the original model and only updating the weights of new layer added on top of the model (Su et al., 2021; Cenikj et al., 2021). In (Cenikj et al., 2021) they use early stopping strategy to prevent overfitting, and trained for 10 epochs or until validation loss of 2 consecutive epochs does not overpass 5*10-3. They show how preprocessing of input texts can improve the task. They evaluate with averaged 10 folds with Macro averaged F1 score.

In (Su et al., 2021), they use the BERT coding of the last hidden layer to train a multi-layer perceptron (MLP). In their experiments, they use BioBERT and PubMedBERT for entity extraction by contrastive learning, a method used in computer vision. In their experiments, PubMedBERT is the best model. However, in (Li et al., 2019a) doing entity normalisation in electronic health records by using BERT models pretrained on biomedical literature or clinical records does not find statistical relevance.

In (Sushil et al., 2021) they explore clinical domain knowledge of BERT and BioBERT, and they show that domain knowledge is challenging even with pretrained language models, because they fail in tasks that require specific domain knowledge. In (Ujiie et al., 2021), for biomedical entity linking, they perform several preprocessing steps to improve performance of PLMs such as splitting documents by sentences, removing punctuations, and resolving abbreviations.

Chapter 3

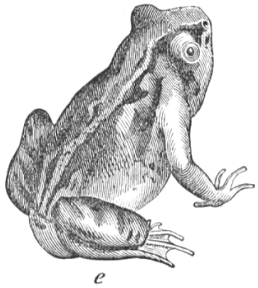
BATRACIO, a dataset for classifying texts in biomedical research phases

This chapter describes in depth BATRACIO, the dataset developed as part of this Master's Thesis.

3.1 Introduction

Batracios, the Spanish translation of "batrachians", are vertebrate amphibians that breathe through their gills when they are born, and through their lungs when they are adults. This is the same idea of evolution throughout the life cycle of these amphibians that remembers the phases of the biomedical research.

Figure 3.1: Picture of a batrachian from *British reptiles and batrachians (1888)* in (Commons, 2020)



The BATRACIO task, acronym of *BAsic-TRAnslational-Clinical research phases classification in bIOmedical publications*, aims to classify biomedical research texts throughout the research phases in the biomedical domain.

Biomedical research involves the application of the natural sciences, especially biology and physiology, to medicine, as the U.S. National Library of Medicine defines[1]. This research area in particular has evolved a lot in recent years, and the number of publications has grown exponentially, but the categories used to classify this area have not evolved in the same way (Laar et al., 2018). Moreover, research in this field is increasingly expensive, and a discovery from its beginning to its end requires the contributions of a broad set of actors carrying out different research activities (Talmar et al., 2020; Mazzucato, 2021). This means that identifying different phases of the same investment is costly and challenging.

Furthermore, in recent years, these areas have become increasingly separated from each other (Butler, 2008) and research actors have become more and more specialised (Talmar et al., 2020). The look towards research from a value-chain perspective (Simpkin et al., 2017) allows to reduce the complexity of research in research phases. The interest for these perspectives has grown in recent years (Hanney et al., 2015; Weber, 2013; Hanney et al., 2015; Flier and Loscalzo, 2017). A valid starting point to grasp the concept of *value chain* is as *"a set of activities that brings a product from its conception to the different phases of production, distribution and consumption, and how these phases and elements that are responsible for each of the phases, coordinate and complement each other"* (Porter, 1985).

Classifying the scientific outputs according to research phases of research funded under a specific funding instruments or scientific publications produced in a region around a specific disease, can help policymakers or funding agencies to better understand what research activities were carried out, mapping stakeholders and their research competencies, and hence to better allocate the resources (Fuster et al., 2020). This information can be used to better steer research funds towards proposals that are more appropriate to the ecosystem or would probably have greater impact, avoiding the risk of allowing funding gaps in different research phases and reducing the risk of duplicating efforts, as (Simpkin et al., 2017) identifies by a manual mapping of funding programmes in UK on antibiotic research.

[1] https://www.ncbi.nlm.nih.gov/mesh?term=Biomedical+Research

The initial goal of this project was to design a task that aims to categorize the documents across different research phases compressed in the biomedical domain. The phases considered by BATRACIO are the following ones: basic research, translational research, clinical research, and public health. For this purpose, it provides an annotated dataset of research publications with the research value chain phase proposed by this project. The first version of the dataset only contains scientific literature. The development of a tool like this, and in combination with other available initiatives, offers a new way of addressing strategic and operational demands of research funding agencies, policymakers, facilitators, and stakeholders, including evidences for the identification and mobilisation of actors, identification of research capacities, evidence-based policymaking, priority-setting and better resource allocation, identify collaboration capacities (inter-institutions and intra-institutional) and new opportunities for monitoring and evaluating impact of research funding.

3.2 Task design

The *BATRACIO* task aims to propose a unified framework for the definition of the phases of the biomedical research value chain and for the allocation of these categories to scientific publications, based on their textual content in title and abstract.

Because there is no agreement between the definitions of the research phases, and the interpretation of the phases may differ depending on the discipline and the perspective of the researcher, the biomedical and health policy experts[2] guiding the project have proposed a new definition and selection of categories, by starting from the existing literature, reviewing publications in detail and emphasising the boundaries between phases. Following the research phases common in cancer research and pharmaceutical development, among others. It has been necessary to have a precise definition of the categories, with practical examples, because there are no clear and universally accepted definitions. Although different attempts to define the phases in biomedical research can be found in literature (Weber, 2013; Hanney et al., 2015; Flier and Loscalzo, 2017; Fort et al., 2017), there are no universally accepted definitions of these phases. A much more complex chal-

[2]From the company SIRIS Academic, https://sirisacademic.com/.

lenge arises in assigning individual publications or research projects to each of the research phases.

As a basis for the evaluation of this task, we have generated a new annotated dataset to classify biomedical literature in the biomedical domain according to the research phases to which the research activity corresponds on a taxonomy of four possible categories. The categories considered in BATRACIO are the following research phases: `basic research`, `translational research`, `clinical research`, and `public health`. Due to our knowledge, there are no similar datasets, so We have built the dataset from scratch with domain experts.

These four categories are presented here, but defined in detail with examples in Appendix C:

- **Basic research**, often called *fundamental research*, focuses on scientific exploration and on building new knowledge, and aims to understand fundamental mechanisms of biology, disease and behaviour. For example, in the case of cancer research, basic research asks how or where mutations occur in DNA and how DNA functions in a healthy cell[3].

- **Translational research**, also called *pre-clinical research* (Woolf, 2008) focuses on translating the discoveries from basic research into usability in the clinic, to produce new drugs, devices and treatment options for patients, with a particular focus on applicability. It uses large-scale testing and both animal models and human biological material, such as computer-assisted simulations of drug, device or diagnostic interactions within living systems.

- **Clinical research**[4] seeks to test a specific treatment or procedure, drug, diagnostic or any technology on patients, focusing not only on the biological mechanisms, but also on issues of safety, delivery and protocols for implementation. It includes studies to better understand a disease in humans and relate this knowledge to findings in cell or animal models.

[3] https://blog.dana-farber.org/insight/2017/12/basic-clinical-translational-\research-whats-difference/

[4] https://blog.dana-farber.org/insight/2017/12/basic-clinical-translational-\research-whats-difference/

- **Public health** phase involves activities to strengthen public health capacities and services seek to provide conditions under which people can stay healthy, improve their health and wellbeing, or prevent the deterioration of their health. For instance, population analyses and retrospective studies, are considered in this phase.

The task is a classification problem, given an input, in our case a scientific publication, the system must identify the most appropriate category according to the content described in the text. For example, given the title of the following scientific publications, a domain expert could assign the following categories:

Category	Title
Basic research	Mechanistic dissection of the PD-L1:B7-1 co-inhibitory immune complex.
Translational research	Optimized fractionated radiotherapy with anti-PD-L1 and anti-TIGIT: a promising new combination.
Clinical research	Infection with multiple hepatitis C virus genotypes detected using commercial tests should be confirmed using next generation sequencing.
Public health	How did the use of psychotropic drugs change during the Great Recession in Portugal? A follow-up to the National Mental Health Survey

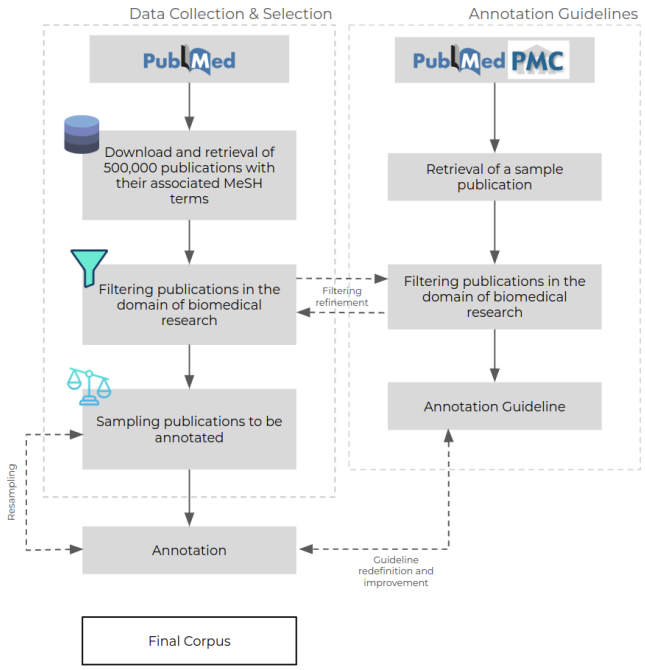
Table 3.1: Example of publications annotated with the four categories.

These publications could undoubtedly belong to these categories, but this manual classification could hardly be done by a person with no knowledge of the biomedical field or who is not familiar with these categories.

The dataset consists of a principal collection of scientific articles which are annotated by three domain experts. The collection includes title and abstract, which are the minimum annotation units to differentiate between the categories proposed. Figure 3.2 presents the outline of the creation of the dataset and the annotation guidelines. The design of the annotation guidelines was done by domain experts, and has allowed us to define the task and the categories, as explained in detail in Section 3.4.

We have ensured a real-world distribution, where classes are not bal-

Figure 3.2: Outline of the process for dataset creation.



anced, to improve the system's generalisability and use in real cases. However, we have avoided biases that could affect the design of the dataset. In this regard and to achieve a representative photograph of biomedical research, we have considered a temporal range that comprises the last five years, from 2015 until 2019, in order to have a certain temporal balance. A longer time frame was ruled out, in order to contemplate the growing trend of translational research. Finally, it was decided not to include the year 2020, due to the high weight of research in COVID-19 that could probably add noise to the dataset. Finally, because expert annotation is costly, we have chosen to annotate 1,000 publications. It is therefore essential to make a selection of the data for the dataset, taking into account possible biases.

The data collection and selection process in the creation of the dataset, and the annotation, are presented in detail below.

3.3 Dataset creation

The first step in the construction of the dataset is to collect a broad sample of publications and projects. In the second step, the records in the biomedical

domain are filtered, because the database contains a broader scope of records in health and life sciences, and for the purpose of the project, we are only interested in the biomedical domain. In the third and final step, we extract a sample of publications, also considering different balancing strategies on a set of features of interest.

The database considered are scientific publications from MEDLINE database, which are extracted from PubMed. PubMed is the search engine of MEDLINE, the bibliographic database on life sciences and biomedicine maintained by the National Library of Medicine (NLM). It covers more than five thousand scientific journals and more than 26 million records. It has a set of APIs that allow retrieval of the publications along with their associated metadata. A large percentage of the records available at PubMed are tagged with the Medical Subject Headings (MeSH terms), mainly assigned by hand by reviewers. These indexing categories will be used for the filtering of articles in the biomedical domain, since the database covers articles beyond the domain of interest.

3.3.1 Gathering publications from PubMed

As we don't have access to the complete PubMed/MEDLINE database, given that the API can only retrieve results from a query, the selection strategy consists of downloading all publication IDs, sampling 500,000 publications and then only downloading data for those records. The reason why we first download all the publication IDs from the API is to ensure representative and balanced sampling when we select the sample to annotate. The APIs allow us to sort the results by date, relevance, citations, but we cannot collect a random set of publications between 2015 and 2019. For this reason, we decide to download all the publications indexed in PubMed for our time range and then filter locally. The download uses two of the main *Entrez Programming Utilities*[5], the public PubMed's API. Firstly, the *ESearch* API[6] allows us to retrieve the list of unique IDs (UIDs) matching the query. The information of each publication is then downloaded by requesting a set of publications from the *EFetch* API[7], which provides description,

[5]Documentation available at https://www.ncbi.nlm.nih.gov/home/develop/api/

[6]API documentation available at https://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.ESearch

[7]API documentation available at https://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.EFetch

title, journal, citations, author, author affiliation, among many other fields[8].

The type of records selected are "journal articles", because this type of record is the majority, with 92.6% of publications indexed between 2015 and 2019. Journal articles tend to be consistent, their abstracts tend to follow similar patterns and structures and have a certain homogeneity in the writing. In the other types of documents indexed in PubMed, the length and structure of the textual content is unknown. Only texts written in English will be obtained.

The number of results for the query[9] between 2015-2019 is 2,511,158 articles in English. We only download the IDs of those publications, and then make a random selection of 100,000 publications per year, 500,000 in total, which are entirely downloaded with their associated MeSH terms (Lipscomb, 2000). For those publications, we store title and abstract, the authors, journal information, and MeSH terms. However, not all records contain all the attributes, and therefore, we will only keep those publications that contain all the required fields, as Figure 3.2 shows.

Step	**# publications**	%
Downloaded publications	500,000	-
With all metadata available & MeSH terms	330,459	0.662
After `Filter 1` - MeSH branches in biomedicine	276,769	0.554
After`Filter 1` & `Filter 2`: Subject areas	145,821	0.292

Table 3.2: Data completeness from initial publication set to filtering in biomedical domain.

3.3.2 Filtering publications in the biomedical domain

Given that the database covers a broader scope than biomedical research, we need to separate what is biomedical research from other fields of relevance to health that are out of our scope. For instance, Health Literacy and Health Education that are in fact Education and Health; or Health manage-

[8] https://www.ncbi.nlm.nih.gov/books/NBK25499/

[9] The query used for each year corresponds to: `("201x/01/01"[Date - Publication] : "201x/12/31"[Date - Publication]) and \journal article"[Publication Type] and english[Language]`

ment which has an impact on clinical research implementation, but is not considered as biomedical research.

Figure 3.3: Main steps in the filtering of publications in the biomedical domain.

A filtering process for separating biomeidcal articles is designed by domain experts based on the Medical Subject Headings (MeSH)(Lipscomb, 2000), in an *ontology-based* fashion. MeSH is a comprehensive controlled vocabulary developed by the National Library of Medicine for searching and indexing biomedical literature (Sayers et al., 2018). Publications in MEDLINE are annotated with MeSH terms which are used in many biomedical text mining applications (You et al., 2020). We identify the MeSH taxonomy as the most suitable to filter articles fast. This filtering process involves an ontological classification based on rules defined by discarding and selecting branches of interest that are associated with biomedical research. This selection of branches on the MeSH taxonomy is done by domain experts. The entire filtering is available in the Appendix A, and an example of this selection is shown in Figure 3.4. The aim of this filtering is to capture the perimeter of the biomedical domain, without being too restrictive, and offering all the topic diversity in the field. In other words, the filtering purpose was to select only publications in the biomedical domain to be annotated by the experts. In terms of evaluation, we were principally interested in precision, because the aim was to obtain a big and representative sample of articles. We do so by defining subsets of the MeSH taxonomy that are aligned with the domain of biomedical research, selection done by our domain experts.

To verify our approach, 100 filtered publications are manually evaluated by domain experts to guarantee the publications offered to the annotators are effectively in the biomedical domain. Since even after this first filter by MeSH branches some publications outside the domain were added, a second filtering step was included in order to remove those publications. To do this filtering by disciplines, we will use the alignment between journals and subjects proposed by Science-Metrix (Archambault et al., 2013). It is an open ontology for classifying scientific journals into bibliometric categories.

Figure 3.4: Sample of MeSH filters.

Scope	Main MeSH Heading	Tree Number / ID	Add	Exclude MESH	Exclude TREE
Biomedical Research	Anatomy	A	TRUE		
Biomedical Research	Diseases	C	TRUE		
Biomedical Research	Chemicals and Drugs	D	TRUE		
Biomedical Research	Analytical, Diagnostic and Therapeutic Techniques, and Equipment	E	TRUE		
Biomedical Research	Health Care	N			TRUE
Biomedical Research	Bioengineering	J01.293.069	TRUE		
Biomedical Research	Biomedical Engineering	J01.293.140	TRUE		
Biomedical Research	Health Occupations	H02			TRUE
Biomedical Research	Statistics as Topic	E05.318			TRUE
Biomedical Research	Models, Theoretical	E05.599			TRUE
Biomedical Research	Imaging, Three-Dimensional	E01.370.350.400			TRUE
Biomedical Research	Equipment Design	D004867		TRUE	
Biomedical Research	Equipment Failure Analysis	D019544		TRUE	
Biomedical Research	Finite Element Analysis	D020342		TRUE	
Biomedical Research	Equipment Safety	D004869		TRUE	
Biomedical Research	Fourier Analysis	D005583		TRUE	
Biomedical Research	Rheology	E05.830			TRUE
Biomedical Research	Games, Experimental	E05.385			TRUE
Biomedical Research	Thermometry	E05.933			TRUE
Biomedical Research	Electrical Equipment and Supplies	E07.305			TRUE
Biomedical Research	Brain-Computer Interfaces	E07.305.076	TRUE		

This approach was used to eliminate those publications that are outside the possible areas of interest. The ISSN and ESSN codes of the journals were used to obtain the bibliometric category of each record. The Science-Metrix taxonomy is composed of 22 research fields, and each field also contains subfields, as Figure 3.5 shows. The complete filters are available in the Appendix B.

Our experts evaluated both filters on a random set of 100 publications again, and the double filtering was found to have a precision of 92.8%. This filtering was applied to the final subset of publications and any publications outside the biomedical research domain found by the annotators during annotation were selected as `Discarded`, as indicated in the Annotation Guideline in Appendix C.

Filter	Precision
`Filter 1`: MeSH branches in biomedicine	78.26%
`Filter 1` + `Filter 2`: Subject areas	92.85%

Table 3.3: Precision of filtering in the biomedical domain on 100 random publications.

Figure 3.5: Sample of Subject filters.

Field_English	SubField_English	REMOVE/KEEP
Agriculture, Fisheries & Forestry	Veterinary Sciences	REMOVE
Agriculture, Fisheries & Forestry	Dairy & Animal Science	REMOVE
Agriculture, Fisheries & Forestry	Food Science	REMOVE
Agriculture, Fisheries & Forestry	Fisheries	REMOVE
Agriculture, Fisheries & Forestry	Agronomy & Agriculture	REMOVE
Biology	Plant Biology & Botany	REMOVE
Biology	Evolutionary Biology	REMOVE
Biology	Entomology	REMOVE
Biology	Zoology	REMOVE
Biology	Ecology	REMOVE
Biology	Marine Biology & Hydrobiology	REMOVE
Biomedical Research	Biochemistry & Molecular Biology	KEEP
Biomedical Research	Microbiology	KEEP
Biomedical Research	Developmental Biology	KEEP
Biomedical Research	Virology	KEEP
Biomedical Research	Toxicology	KEEP
Biomedical Research	Physiology	KEEP
Biomedical Research	Nutrition & Dietetics	KEEP
Biomedical Research	Genetics & Heredity	KEEP
Biomedical Research	Biophysics	KEEP
Biomedical Research	Mycology & Parasitology	KEEP
Biomedical Research	Anatomy & Morphology	KEEP
Biomedical Research	Microscopy	KEEP
Built Environment & Design	Building & Construction	REMOVE

3.3.3 Selection of publications

After applying the filtering in the biomedical domain, we have 145,821 publications from which we have to select 1,100 candidate publications to be annotated. As described above, this sampling should be representative and balanced according to different factors. In order to ensure this and to avoid bias, we sample according to the following features of interest:

- Publication Year
- Affiliation Country of the Authors
- Journal
- Anatomy (MeSH branch A)
- Organism (MeSH branch B)
- Diseases (MeSH branch C)
- Chemicals and Drugs (MeSH branch D)

Maintaining an appropriate distribution in the A, B, C and D branches of the MeSH taxonomy will allow for generalizability of the dataset, because

a very large sample of a specific disease in any of the annotated categories could add noise and bias the dataset, and would make it difficult to identify generalizable patterns in the texts.

3.4 Annotation

A manual annotation process is preferred, since especially in the medical domain, the quality of the annotations is more important than the quantity of annotations (Patel et al., 2018). For the dataset annotation process, we have been inspired and oriented by the corpus creation methodologies proposed by (Szarvas et al., 2008; Patel et al., 2018; Oconnor et al., 2020), for the biomedical domain, which include the following steps:

1. Annotation guide and study of sample examples to create the guide.
2. Training of the annotators.
   (a) Labelling a sample of texts.
   (b) Discussion about conflicts.
   (c) Updating the annotation guide to cover conflicts.
3. Annotation of the dataset.
   (a) Annotation of the 20% of the records.
   (b) Agreement at first 20%.
   (c) Redefinition of the annotation guideline.
   (d) Annotation of the whole dataset.
   (e) Final agreement.
4. Dataset statistics and description.

3.4.1 Guideline development

Since this is a new task, the annotation guideline are designed from the ground up. For the definition, two experts in the domain have defined the boundaries of the categories based on literature review and expert discussion. In the initial project proposal, after reviewing the main phases in biomedical research in literature, we only considered the three categories of basic research, translational research and clinical research, assuming that they were

able to cover the whole spectrum of biomedical research. To adapt these preliminary definitions to an annotation scenario, both experts attempted to annotate 100 publications in the biomedical domain (randomly extracted from PubMed) with the three proposed categories. After this process, they expressed the need to add a phase after clinical research, to cover the whole spectrum of publications considered as biomedical research, because some of them are focused on socioeconomic aspects, retrospective studies about impact on population, and related to health policy issues of biomedical research. We have named this fourth phase "Public Health", although it is idfferent from the general concept of public health, because we are here restricted to biomedical research. After this, the guidelines are adapted and complemented with a set of examples for each of the categories.

In the same experiment, the minimum unit of annotation is evaluated. Given the suspicion that only the title and the abstract would not be sufficient for some articles, the sections of introduction, materials and methods, and MeSH terms, are also provided to the experts for each article. Unexpectedly, the experts conclude that by using just the title and abstract, it is possible to assign all four categories to scientific publications. This is an interesting finding for future applications of this resource, since 52% of scientific articles in the fields of life sciences and molecular biology are not open access and only the title and description are accessible[10].

The differentiation between some of the phases can be enormously challenging. In some cases, the category suitability can be clear, because the methods and type of research activity are undoubtedly within a specific category. However, in others the difference could become difficult to distinguish, because what differentiates the phase is the scientific question behind the research activity. Furthermore, in other cases, the document can be near the borders between two areas, because the task forces the documents to belong to one class. To help annotators in ambiguous cases, the categories have been defined exhaustively by experts. The definition has also benefited from the discussion about a sample of documents.

After the annotation of the first 400 publications, guidelines were updated by redefining boundaries between categories, because between basic and translational research the agreement was specially low, a slight agreement of 0.22. This made it necessary to adjust the annotation guidelines,

[10]PubMed

because some publications were difficult to interpret or did not meet the initial definition of the guidelines.

3.4.2 Annotation Procedure

The 1,100 publications were annotated by 3 expert in the domain. The three annotator have a PhD in different fields of biomedicine and developmental biology. Agreement was calculated on the base of averaged Cohen's k, and the three annotators annotated the same publications, due to the complexity of the task.

Firstly, we annotated the first 400 publications, where the general agreement was moderate. When studying the agreement between pairs of categories, between basic and translational research the agreement was specially low, a slight kappa agreement of 0.22. After a discussion between the annotators and a review over some publications with zero and partial agreement, the first 400 publications were re-annotated. The agreement increased substantially our kappa to 0.78 (+.09) with respect to the first annotation. This agreement is particularly high, and the definitions are considered as good. During the following iterations, the quality is maintained, even between pairs of categories, as Table 3.4 and Table 3.5 show. Finally, the 14 publications with zero agreement were discussed, and the final label was agreed among the three annotators based on the definition in the annotation guidelines.

Table 3.4 shows the results of the inter-annotator agreement by annotation checkpoint and iteration, as well as the final *kappa*-score. The final agreement is adequate between the three annotators. However, as more publications are annotated and more time passes since the initial discussions, the agreement decreases slightly. In fact, the accuracy of manual text classification can be influenced by human factors such as fatigue and expertise (Li et al., 2020). On the other hand, Table 3.5 shows the agreement by pairs of categories, and one can observe that the most difficulty is found between pairs of categories that are adjacent as research phases, such as basic-translational, translational-clinical, or clinical-public health. In addition, after the iteration 2 on the checkpoint 1, the boundaries of each category are redefined. However, this differentiation remains particularly challenging, given the high overall inter-annotator agreement.

Two control checkpoints were selected in order to explore general agreement and by pairs of categories, and also to explore the number of instances

	Cohen's k	% Coincidence
Annotation of first 400 publications	0.690	0.730
Re-annotation of first 400 publications, after redefinition of guidelines	0.782	0.859
Annotation of first 800 publications	0.768	0.850
Final dataset (1100 publications)	0.748	0.832

Table 3.4: Average of inter-annotator agreement between the 3 annotators during the development of the annotation guidelines, during the annotation of the dataset, and for the final dataset.

of each category.

The annotation was carried out by three annotators, the two annotators who designed the guidelines and a third annotator. A1, A2 and A3 annotated the same publications. In general, the inter-annotator agreement is substantial. In the first annotation checkpoint, on the first 400 publications, a k=0.69 was obtained, which is acceptable. Although when looking at the agreement by pairs of classes, between the Basic-Translational and Clinical-Public Health categories, this agreement is particularly low. After this, the categories were redefined and some of the cases of disagreement were discussed in order to reformulate guidelines. A second annotation iteration on the same 400 publications was carried out, and it achieved a substantial agreement for all pairs of categories. At checkpoint 2 after the first 800 publications, the agreement remained acceptable, but given the proportion of publications in the Basic and Translational categories with respect to the clinical research, a half of the remaining sample of 300 publications to be annotated is resampled based on a selection of journals which could be more related with basic and clinical research.

3.5 Dataset Statistics

In Table 3.6, we show general class distribution. Given the small size of the Public Health category, we show both distribution of classes, also removing this class. Among the 1100 publications, 480 (43.64%) are in the clinical research phase, 349 (31.73%) in the basic research phase, 220 (20.00%) in the translational research phase, and 51 (4.64%) in the public health phase.

Table 3.7 shows the general characteristics of the dataset. There are

Cohen's k			
Pairs of categories	Annotation of first 400 publications	Re-annotation of first 400 publications, after redefinition of guidelines 1	Final dataset
Basic-Translational	0.223	0.614	0.595
Translational-Clinical	0.796	0.904	0.867
Clinical-Public Health	0.617	0.719	0.719
Basic-Clinical	0.887	0.953	0.946
Basic-Public Health	1.000	1.000	0.962
Translational-Public Health	0.986	1.000	0.994

Table 3.5: Average of Cohen's k inter-annotator agreement between the 3 annotators during the development of the annotation guidelines, during the annotation of the dataset, and for the final dataset.

1100 publications in the biomedical domain annotated in 4 categories. The total number of sentences in the dataset is 11,066, in only title and abstract, with an average of 10.06 sentences per publication. The total number of uncased words in the dataset is 278,103, with 25,834 unique words, and the average number of words per publication is 252.82.

	ALL Classes	Removing Public Health
Basic research	349 (31.73%)	349 (33.27%)
Translational research	220 (20.00%)	220 (20.97%)
Clinical research	480 (43.64%)	480 (45.76%)
Public Health	51 (4.64%)	-
Total	**1100**	**1049**

Table 3.6: Class distribution in the final dataset.

	Final dataset
#categories	4
#documents	1,100
#sentences	11,066
avg. #sentences	10.06
#words (total)	278,103
#words (unique)	25,834
avg. #words	252.82

Table 3.7: Final dataset statistics.

Chapter 4

Systems and classification methods

The aim of this work is not only to provide datasets, but also to see if we can automatically classify biomedical scientific publications according to the BATRACIO, i.e. according to four categories which represent the different phase of the biomedical research. For this purpose we have used and evaluated a wide range of current state-of-the-art systems, which we have grouped into traditional machine-learning systems, deep-learning systems, and BERT-based systems.

4.1 Traditional machine-learning methods

Most *traditional* machine-learning methods follow a two-step approach, based on feature extraction, which is fundamental for the effectiveness of the method, and then to feed a classifier with those features. There can be different types of features, which generally are split sentences, preprocessing texts, cleaned and segmented words, converted to vectors by filtering words by frequency, or other weighting algorithms (Kalyan et al., 2022). For this case, we have tokenized words and removed stopwords, then we extract features from text by TF-IDF, the number of features extracted are 10,000, because this kind of algorithms do not work well with large numbers of features.

The learning methods used in our experiments are SVM and Random Forest, two of the machine-learning algorithms most popular in text classi-

fication (Kowsari et al., 2019). These systems have been implemented with the library `Sklearn`[1].

4.1.1 SVM

SVM (Sebastiani, 2001) is one of the most popular methods in text categorisation due to its good performance. It learns the hyperplane in a feature space that separates the positive training examples from the negative ones with the maximum possible margin, tending to minimize the generalization error. This learning algorithm was originally designed for working with binary classification tasks, and the dominate technique for addressing multi-class problems with SVM is by using Multiple-SVM (Kowsari et al., 2019).

- **SVM-Tfidf**: One of our baseline models is based on a SVM classifier with default parameter setting in SKlearn, This means that a linear kernel has been used. Features are extracted by applying TF-IDF on preprocessed text and top 10,000 features are selected. We empirically chose this number because this kind of algorithms do not work well with large numbers of features.

4.1.2 Random Forest

Random Forest (RF) (Kowsari et al., 2019) is a learning method based on a combination of decision trees, the predicted class is the category selected by most trees. RF and decision tree methods are fast and accurate for document categorization. They are fast to train but slow in prediction steps.

- **RF-Tfidf**: One of our baseline models is based on a Random Forest classifier with default parameter setting in SKlearn. This means that the classifiers has 100 trees. Features are extracted by applying TF-IDF on preprocessed text and top 10,000 features are selected. We empirically chose this number because this kind of algorithms do not work well with large numbers of features.

[1] https://scikit-learn.org/stable/

4.2 Deep learning methods

In the last 10 years, deep learning-based techniques have improved traditional shallow machine learning, by learning a set of nonlinear transformations that allow to map features directly to output, including feature extraction into the model fitting process (Li et al., 2022). These methods can be trained on unstructured data and can learn feature representation directly from input text without too many manual interventions (Li et al., 2022). The learning methods used in our experiments are CNNs and LSTMs, two of the most used architectures in NLP.

These systems have been implemented with the library `Keras`[2], and for the configuration of them we have been inspired from (Bokka et al., 2019).

4.2.1 Convolutional Neural Networks (CNN)

A convolutional neural network (CNN) (Kowsari et al., 2019) is a deep learning architecture that, although initially designed for images, has been widely used for text classification. CNN can automatically extract features from texts, applying convolutional filters of different sizes. These systems capture well local information, but have difficulties to capture information at long distance in the text. Pre-trained word embeddings improve generalizability and allow applying transfer learning in word representation. However, general-domain pre-trained word embeddings lack of sufficient knowledge of biomedical domain, for this reason we explore the use of general-domain Word2vec (Mikolov et al., 2013), pre-trained on Google News, and BioWordVec (Yijia et al., 2019), a pre-trained Word2vec model in biomedical literature.

- **CNN-Word2Vec**: A model with pre-trained vectors from `Word2Vec`, batch size of 16, one convolution layer with kernel of 5, and global max pooling.

- **CNN-BioWordVec**: A model with pre-trained vectors from `BioWordVec`, batch size of 16, one convolution layer with kernel of 5, and global max pooling.

[2] https://keras.io/

4.2.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) are a type of Recurrent Neural Network (RNN) that consider the problem of keeping long distance information and dependency, unlike CNN. They learn sequential information by computing sequentially. They appear to solve the vanishing gradient problem by introducing memory to the network. In our experiment we have used Bidirectional LSTM (biLSTM), which is a sequence processing model that consists of two concatenated LSTM, the first takes the input in a forward direction and the other, backwards. They increase the information available about context, although this is not exactly bidirectional. However, the main handicap of this model is that it is are sequential and time-consuming.

- **biLSTM-Word2vec**: A model with pre-trained vectors from `Word2Vec`, batch size of 16, and one layer with 32 units.
- **biLSTM-BioWordVec**: A model with pre-trained vectors from `BioWordVec`, batch size of 16, and one layer with 32 units.

4.3 BERT-based systems

In the last few years, pre-trained language models based on Transformers have stood out as a better option for text classification, due to their capacity of learning global language representations from large datasets and being adapted to a downstream task just by adding a final layer, in the case of text classification a linear classifier, and adapting the weights of the model by fine-tuning parameters. Specifically, BERT is the first pre-trained fine-tuning based and bidirectional language model, that achieve state-of-the-art results in several NLP tasks. Contextual pre-trained language models improve previous models because they do not have to be trained from scratch, reducing cost of computational resources and improving in performance.

Specifically in BioNLP tasks, BERT-based systems are the new baseline method, as we can see by reviewing contribution in last three years of BioNLP workshop in ACL. In most of the contributions in BioNLP, in-domain pre-trained language models tend to improve versus general pre-trained models. This is because the domain language is specially challenging and complex, although in others specific domains, like financial domain, there is not such a clear improvement with domain-specific models (Arslan

et al., 2021). In recent years, new approaches for adapting the pre-training language models have emerged improving the biomedical knowledge of them have appeared. We have explored the use of some of the most widely used BERT-based biomedical models and some other interesting although less used in literature.

The implementation of the systems is carried out by using HuggingFace's Transformers library as the pre-trained language models.

There are many models available, and to identify which model is the best for a specific task can be a challenge, given the increasing number of pre-trained language models and architectures. This step implies an intensive use of computational resources (Nozza et al., 2020).

The following systems have been implemented with `Pytorch`[3] and models have obtained from `HuggingFace`[4].

4.3.1 Pre-trained BERT models and biomedical variants used

The models used, and their main features, are described below:

- **BERT-base (Devlin et al., 2019)** is a multi-layer bidirectional Transformer encoder. Pre-trained on general domain corpus, BooksCorpus and English Wikipedia, for the objectives of Masked Language Modelling and Next Sentence Prediction. We chose BERT-base architecture (12 layers, 768 hidden learning and 12 attention heads, summing a total number of 110M parameters), for the comparison with biomedical variants of BERT.

- **BioBERT (Lee et al., 2019)** is initialised from weights of general-domain BERT (Devlin et al., 2019), and it is further pre-trained on PubMed abstracts and PubMedCentral full-text articles. They keep the same vocabulary of general-domain BERT.

- **SciBERT (Beltagy et al., 2019)** is a BERT-base model adapted to the specific-domain by pre-training on a random sample of mixed-domain 1.14M full text papers from Semantic Scholar, 18% in computer science and 82% in the biomedical domain. It includes new vocabulary in scientific domain which only overlaps 42% with the general-domain vocabulary in BERT and BioBERT.

[3] https://pytorch.org/
[4] https://huggingface.co/

- **PubMedBERT (Gu et al., 2022)** is a domain-specific BERT-base model pre-trained from scratch on PubMed literature.
- **BlueBERT (Peng et al., 2019)** starts from weights of general-domain BERT, and it is further pre-trained on PubMed abstracts and clinical notes, demonstrating the importance of pre-training among different text genres for some tasks.
- **OuBioBERT (Wada et al., 2021)** is pre-trained on general-domain documents and a small corpus of biomedical abstracts from PubMed. It applies simultaneous pre-training on small dataset in the medical domain with general-domain documents, based on a well-balanced pre-training by up-sampling instances derived from a corpus appropriate for the target task, and also with an amplified vocabulary, in order to face the problem of specific domains with small number of available documents as well as specific domains in other languages.
- **BioLinkBERT (Yasunaga et al., 2022)** is pretrained in the same pretraining corpus used in PubMedBERT, PubMed abstracts, adding citation between articles as additional information during pretraining, in the new Document Relation Prediction pre-training objective task.
- **ClinicalBERT (Alsentzer et al., 2019)** is initialised from BERT-base and further pre-trained over approximately 2M million clinical notes.

4.3.2 Adaptation of BERT models

We have explored the two main approaches for adapting BERT models to a text classification task (Tunstall et al., 2022). The first one, fine-tuning, proposed in the original paper (Devlin et al., 2019) and other studies (Sun et al., 2019; Su et al., 2021; Cenikj et al., 2021). Fine-tuning consists in adjusting and updating pre-trained weights by a back propagation algorithm that aims to reduce loss function value in order to get closer to desired outputs. The second, feature-based approach, consists in all parameter freezing and simply training a linear classifier embedded on top of the model, as suggested in (Cohan et al., 2020; Pascual et al., 2021). Furthermore, parameter freezing could be a good approach for avoiding catastrophic forgetting (McCloskey and Cohen, 1989; Sun et al., 2019), because our dataset is small.

Following the proposed optimal hyperparameters proposed in the original BERT paper (Devlin et al., 2019), we have explored the following hyperparameter configuration:

- **Learning rate (Adam):** 5e-5, 3e-5, 2e-5
- **Batch size:** 16, 32
- **Epochs:** 2, 3, 4

4.3.3 Loss function

The main objective of neural networks is to minimise the difference between the prediction and the expected output, comparing the predicted distribution of results and the true distribution. This distance or error is computed based on a cost or error function. The standard cost function for text classification tasks is the *cross-entropy loss*[5].

However, the cost function does not consider the class imbalance. Cost weighting is an important alternative to data augmentation for unbalanced classes (Tayyar Madabushi et al., 2019), it involves increasing the cost associated with obtaining an erroneous low-frequency class label. A special feature of our task is that categories are sequentially sorted. Nevertheless, cost function does not consider order between categories. We propose an ordered weighting of the loss function by considering the error to be double when categories, target and predicted, are adjacent categories, because this is where the highest error rate is concentrated.

In our experiments, we consider the following modification in the loss function:

- **Loss**: cross-entropy loss without weighting categories.
- **Weighted loss**: as (Tayyar Madabushi et al., 2019) propose, we increase the cost of incorrectly labelling the class with lower number of samples by weighting the cross-entropy loss function.
- **Neighbouring loss**: we consider the error, from cross-entropy loss, to be double when target and predicted categories are adjacent.

[5] https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

4.3.4 Text preprocessing

One of the strengths of BERT is that systems can directly learn from unstructured text; however, in the biomedical domain, learning from unstructured language remains a challenge (Schick and Schütze, 2019; Pascual et al., 2021; Cenikj et al., 2021). Acronyms are especially common in science and even more in biomedical publications, as authors regularly seek to shorten the long names for diseases, bacteria, and chemicals. Barnett and Doubleday (2020) documented acronyms use in more than 24 million scientific article titles and 18 million scientific articles published between 1950 and 2019. They reported that 19% of titles and 73% of abstracts contain acronyms. Of the more than one million unique acronyms in their data, 0.2% appeared regularly and most acronyms, 79%, appeared less than 10 times (Hogan et al., 2021).

Schick and Schütze (2019) says that pre-trained language models do not work well with rare words, and when datasets have a high number of unique words, they are also more challenging. In the biomedical domain, several articles propose to pre-process and clean documents to improve performance of pre-trained language models. Pascual et al. (2021) propose to remove all numbers, because they are frequent in scientific studies, but they do not add relevant information for BioNLP tasks. Cenikj et al. (2021) also highlight how preprocessing can improve performance tasks for the same reasons. In (Ujiie et al., 2021) punctuation and abbreviations are removed.

We perform basic processing following the recommendations in (Schick and Schütze, 2019; Cenikj et al., 2021; Pascual et al., 2021; Ujiie et al., 2021):

- **Acronym resolution:** we use the `AbbreviationDetector` component in `scispacy`[6], which implements a simple algorithm for identifying abbreviation's definitions in biomedical text (Schwartz and Hearst, 2003) and after, the abbreviations are replaced by the full name.
- **Removal of numbers:** we remove all numbers in abstracts, they do not add meaning about the categories of interest, and sometimes correspond to results or references.
- **Removal of special characters:** we remove special characters, because scientific literature can include formulas and rare characters, which can reduce performance.

[6] https://github.com/allenai/scispacy

Chapter 5

Evaluation and discussion

This chapter describes the methodology used to evaluate the proposed systems and methods, and presents the results obtained in different sets of experiments. The results are also analysed and discussed in depth.

5.1 Evaluation methodology

5.1.1 Evaluation collection

The dataset we have used is BATRACIO, a dataset designed and developed as part of this project, and explained in detail in Chapter 3. BATRACIO consists of 1100 scientific publications from PubMed with title and abstract. The publications have been annotated by three domain experts with the following categories: `basic research`, `translational research`, `clinical research`, and `public health`. These categories form one of the main understandings of the phases of biomedical research. Biomedical research involves different phases, because a complete discovery often requires the involvement of many actors doing different things.

The dataset is imbalanced, and Table 5.1 presents the distribution of categories in the dataset. Table 5.2 describes some of the main characteristics of the dataset, such as the number of sentences and the numbers of words. For more information on the dataset, see Chapter 3.

5.1.2 Evaluation metrics

We use `precision`, `recall`, `f-measure`, and `accuracy`, measures in order to evaluate the behaviour of SOTA machine learning techniques. The eval-

	ALL Classes	Removing Public Health
Basic research	349 (31.73%)	349 (33.27%)
Translational research	220 (20.00%)	220 (20.97%)
Clinical research	480 (43.64%)	480 (45.76%)
Public Health	51 (4.64%)	-
Total	**1100**	**1049**

Table 5.1: Class distribution in the final dataset.

	Final dataset
#categories	4
#documents	1,100
#sentences	11,066
avg. #sentences	10.06
#words (total)	278,103
#words (unique)	25,834
avg. #words	252.82

Table 5.2: Final dataset statistics.

uation of the systems by means of the dataset has been carried out through 10-fold cross-validation. The split is done in a stratified approach to ensure samples of all categories in all partitions, as random partitioning did not provide such representation. The validation set is carried out by a 0.1 split on the train. As we will evaluate with k-fold cross-validation, we use the averaged values of the metrics across all k-folds. The evaluation of imbalanced datasets is challenging, because models generally predict the majority class with major accuracy, and because of this, results can be misleading. Macro-averaged F-measure is a more appropriate evaluation metric, as it considers the performance of each class equally (Tayyar Madabushi et al., 2019).

The metrics are described in detail below:

- **Precision**[1] is the proportion of correctly classified examples among those classified in that category. It is a good metric to identify how well it identifies the samples predicted in the category, but it does not include information about the quantity of results. This is calculated as:

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html#sklearn.metrics.precision_score

$$Precision = \frac{\#True\ positives}{\#True\ positives + \#False\ positivies}$$

- **Recall**[2] is the proportion of true categories that are identified. This is calculated as:

$$Recall = \frac{\#True\ positives}{\#True\ positives + \#False\ negatives}$$

- **F-measure** combines precision and recall, by calculating the harmonic mean between both. Precision and recall are complementary metrics, and f-measure can combine information from both. This is calculated as:

$$F\text{-}measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- **Accuracy** is the share of correct predictions (both true positives and true negatives) among the total number of cases examined. However, it presents some limitations when dealing with imbalanced datasets, as a non-informative system that classifies all instances as the majority class could achieve considerable good accuracy results, by counterpart with f-measure. This is calculated as:

$$Accuracy = \frac{\#Correct\ classifications}{\#All\ classifications}$$

We choose these evaluation metrics because they are the most widely used in the literature, they allow us to compare with other works, and because they are best suited to the evaluation of a categorical classification system.

The metrics are calculated with the library `sklearn`[3].

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html#sklearn.metrics.recall_score

[3] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

5.2 Results

The experiments performed in this section focus on exploring the use of domain-specific pre-trained language models to solve the BATRACIO task, and how to adapt them to a downstream task of text classification in order to obtain satisfactory results.

The following research questions have been addressed in the experiments:

1. Can the problem be addressed automatically with general-domain systems? Can recent pre-trained language models improve on traditional methods?
2. Do domain specific pre-trained language models outperform generic pre-trained language models?
3. Which is the best approach to adapt models to the text classification task, the fine-tuning or the feature-based approach?
4. What input data do we need to automatically address the problem?
5. Does the imbalance of the dataset affect the performance of systems? Can we mitigate this fact by forcing the system to learn from class distribution and semantic distance between categories?
6. Do pre-trained language models improve if text is pre-processed?
7. How important are hyperparameters? And what would be a pertinent configuration?

5.2.1 Can the problem be addressed automatically with general-domain systems? Can recent pre-trained language models improve on traditional methods?

Table 5.3 shows the results of the general-domain systems considered as a baseline, using as input title and abstract. Among these, we have explored traditional machine learning systems on linguistic features, and deep-learning-based systems using general and domain-specific word embeddings. The best result is obtained with BERT-base model adapted to the task with fine-tuning. As it can be seen, all other systems are not able to perform the task adequately, as they only predict the two majority categories. The fact that systems that typically perform well in many text classification tasks do

not satisfactorily solve this task may indicate that the task is complex and challenging.

System	F1	Acc.
SVM-tfidf	.511	.767
RF-tfidf	.499	.756
CNN-Word2Vec	.470	.638
CNN-BioWordVec	.427	.645
biLSTM-Word2Vec	.479	.680
biLSTM-BioWordVec	409	.632
BERT-base		
- *Fine-tuning*	**.785**	**.871**
- *Feature based*	.245	.462

Table 5.3: Comparison of an ensemble of general-domain methods used in text classification. Input data is title+abstract. Results are reported as macro-average between 10-folds.

Some of the common issues between the systems is that extraction of features by weighted frequency from documents does not work correctly for those systems that learn from features, because the link to the category is not explicitly mentioned, as would be the case if the categories were diseases. It is clear that linguistic and semantic features present difficulties representing categories in BATRACIO, including non-contextual word embeddings[4] and contextual embeddings[5], because they are encoding the same linguistic and semantic information of the documents, where some concepts related to diseases, molecular entities or viruses, which are transversal, can take special relevance in the vector representation. Of course, an accurate design of features, involving domain experts, could improve the performance of baseline systems.

As extracted from the results, the use of word embeddings, general-domain and specific-domain, does not provide much information to help systems to clearly differentiate between categories. However, fine-tuning all parameters with BERT-base seems to solve the task much better than previous systems. According to our domain experts, the difference between categories is essentially based on the research question behind the scientific article, and often this is not explicitly mentioned in the abstract, and strong domain knowledge is needed to contextualise this information from the title

[4] As Word2Vec and BioWordVec.
[5] Alluding to the feature-based implementation of BERT.

and abstract. The fact that it is an imbalanced dataset makes it even more difficult to learn the categories when there are no clear linguistic features. We conclude that BERT-base is the best system explored. Because with its by-default design for text classification (Devlin et al., 2019), it is able to predict all categories and obtain the best evaluation results compared to the other systems explored. However, freezing all parameters of models to simply train a linear classifier on the contextual embeddings does not work in this case, in contrast to what is indicated in (Devlin et al., 2019; Peters et al., 2019).

5.2.2 Do domain specific pre-trained language models outperform generic pre-trained language models on a dataset for multi-class classification?

In Table 5.4, we present experiments on several biomedical BERT-based models, which are described in Section 4.3.1, in order to explore one of the main questions of this project, namely if domain-specific pre-trained language models improve general-domain models, *which kind of domain adaptation can improve performance in this task?* As it can be seen in the table, higher performance is achieved if we use domain-specific models for the biomedical domain, as biomedical domain contains many peculiarities such as specific terminology, polysemic words, and frequent use of abbreviations and acronyms. Indeed, several models pre-trained in the domain or adapted to the domain improve BERT-base, which is designed for the general domain, but not all domain-specific models perform well.

The two best performing models are PubMedBERT and sciBERT, both of which have been trained from scratch on biomedical literature. However, PubMedBERT improves by 2.9% points to sciBERT in F-measure, maybe because PubMedBERT is fully-trained on biomedical documents, but sciBERT is pretrained on 18% of articles from the computer science domain and 82% from the broad biomedical domain. This is particularly relevant to see how pre-training influences the performance of pre-trained language models in specific domains. BioBERT, which shares the vocabulary of BERT-base and was further pre-trained on biomedical documents, does not respond so well, maybe because it does not have vocabulary in the domain and it cannot encode very particular scientific jargon. Specter is the third best, with a difference of 3.7—% less F-measure than the best system. It is presented as a

System	P	R	F1	Acc.
PubMedBERT	**.888**	**.837**	**.850**	**.907**
sciBERT	.851	.809	.821	.821
Specter	.844	.803	.813	.883
BioBERT	.870	.793	.809	.888
ouBioBERT	.860	.781	.799	.883
linkBioBERT	.851	.779	.796	.885
BERT-base	.834	.766	.785	.878
clinicalBERT	.861	.759	.780	.859
BlueBERT	.815	.760	.776	.869

Table 5.4: Comparison of domain specific pre-trained language models. Trained by fine-tuning, fixing the following hyperameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Input data is title+abstract. Results are reported as macro-average between 10-folds.

model that works especially well for classifying documents in research topics. Since including citation information between documents during pre-training could help capture our categories where articles from the same phase could be cited more frequently. It is noteworthy BERT-base's success, it obtains similar results to several other models in the domain, despite being a generic model. One of the possible reasons, as suggested by (Sushil et al., 2021), is that BERT-base was pre-trained on Wikipedia which includes WikiMed, a set of Wikipedia medical and scientific pages, and therefore has some knowledge of the domain, unlike general-domain Word2vec (Mikolov et al., 2013), which is pre-trained on Google News. Surprisingly, BlueBERT which is a model that achieves state-of-the-art results in several BioNLP tasks, has obtained worse results than BERT-base with this hyperparameter configuration. Nevertheless, as we have observed in other experiments not shown in the table, with a by-default learning rate of 2e-5 this model obtains worse results than with a smaller learning rate such as 3e-5 and 5e-5. This points out one of the main challenges when adapting PLMs to a downstream task as, especially when the dataset is small, fine-tuning parameters with high learning rates could remove the knowledge embedded in the model.

5.2.3 Which is the best approach to adapt models to the text classification task, the fine-tuning or the feature-based approach?

Table 5.5 shows the comparison between the adaptation of pre-trained language models to the text classification task by the fine-tuning and the feature-based by approaches. In fine-tuning, it adapts the parameters for minimising the loss between categories. The latter consists of freezing the model parameters and only training the classification layer. BERT is pre-trained on two tasks, Next Sentence Prediction and Masked Language Modelling, based on the knowledge captured during these two tasks, the model is able to generalise by language in a low dimensional space. Basically, fine-tuning consists of adding a linear classifier on top of the last BERT layer and then training the entire network during few epochs, updating all parameter weights by the loss function and training. The feature-based approach consists of only training the classification layer on top of the last hidden state, which provides sentences encoded in 768 dimensions. As can be extracted from the results, feature-based approach does not perform the classification satisfactorily, probably because contextual embeddings can not encode information about the categories of the task, which are transversal across diseases, domains, and techniques, because the information it attends to could not be the most relevant to our categories.

	Feature based		**Fine-tuning**	
System	**F1**	**Acc.**	**F1**	**Acc.**
PubMedBERT	.201	.333	**.850**	**.907**
sciBERT	.168	.431	.821	.821
Specter	.147	.405	.813	.883
BioBERT	.153	.432	.809	.888
ouBioBERT	.151	.435	.799	.883
linkBioBERT	.264	.477	.796	.885
BERT-base	.245	.485	.785	.878
clinicalBERT	.155	.438	.780	.859
BlueBERT	.231	.462	.776	.869

Table 5.5: Comparison of domain specific pre-trained language models by fine-tuning and feature-based approaches. Fixing the following hyperparameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Input data is title+abstract. Results are reported as macro-average between 10-folds.

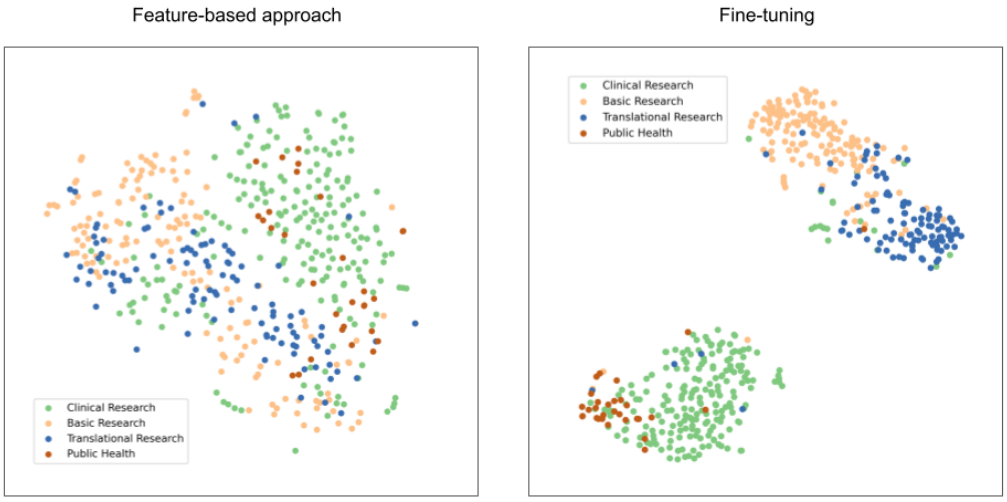


Figure 5.1: T-SNE dimensional reduction of embedding produced by PubMedBERT by feature-based approach (left) and fine-tuning (right). 50% of the dataset was used for fine-tuning the model and the other 50% publications have been encoded with both methods.

With the aim of exploring empirically how the two methods represent the documents, we employed dimensional reduction with T-SNE to visualise the final embeddings of the PubMedBERT model, our best pre-trained language model. For this specific exploration, we have split the dataset in two, using 50% for testing and 50% to train. The image on the left in Figure 5.1 shows the 2-dimensional distribution of the embedding coding that comes out of the PubMedBERT model, and this is used to train a classifier. We have used the other 50% of publications to obtain the encoded contextual embedding produced by the raw model and the fine-tuned version. Indeed, we see how fine-tuning guides the model to clearly differentiate the categories as can be seen in the Figure 5.1, not only for the training and validation set, but also for a set not seen during training.

5.2.4 What is the most important part of the text to automatically address the problem?

The following Table 5.6 displays the evaluation of the models by fine-tuning according to different data inputs. For this purpose, different data input combination were tested (i) title, (ii) abstract and (iii) title and abstract. The combination of both sections, title and abstract, gives better results for all four systems. This experiment is only carried out on the best systems and the most widely used in the literature, using the best configuration. However, it is interesting to see how with only the title the systems can learn to differentiate between the categories. Therefore, both sections are relevant to solve the task correctly. The best system given (ii) abstract, and (iii) title and abstract, is PubMedBERT. However, with the title only, the difference between systems is much smaller. Obviously, (iii) title and abstract contains more information, which allows a better contextualisation of the activities carried out in each publication.

The two sections were described as the minimum unit of annotation by our domain experts, and both may contain relevant information. Although most information is concentrated in the abstract, because there are defined the research problem and contribution of the article, the title often contains very explicit references to the categories of interest such as this publication *Clinical Characteristics of Rheumatoid Arthritis Patients Achieving Functional Remission after Six Months of Non-tumor Necrosis Factor Biological Disease-Modifying Antirheumatic Drugs (DMARDs) Treatment* in the "Clinical Research" phase.

	Title		**Abstract**		**Title+Abstract**	
System	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**
PubMedBERT	.794	.855	.813	.888	**.850**	**.907**
sciBERT	.795	.864	.804	.880	.821	.887
BioBERT	.786	.858	.794	.876	.809	.888
BERT-base	.705	.804	.772	.869	.785	.878

Table 5.6: Comparison of domain specific pre-trained language models trained on different textual sections of the scientific publications. Trained by fine-tuning, fixing the following hyperparameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Results are reported as macro-average between 10-folds.

5.2.5 Analysing the categories in depth

One of the main challenges of working with small and imbalanced datasets is that the systems tend to learn the most frequent classes. Table 5.7 shows F-measure and accuracy by category for PubMedBERT, sciBERT, bioBERT, and BERT-base, which are four main systems in literature and with the highest results in our task. As it can be seen, the best model is PubMedBERT which is also the one that best predicts the categories with fewer instances, namely Public health and Translational research.

	Basic	**Translational**	**Clinical**	**Public Health**	**ALL**
System	**F1**	**F1**	**F1**	**F1**	**F1**
PubMedBERT	**.912**	**.850**	**.949**	**.688**	**.850**
sciBERT	.892	.816	.937	.639	.821
BioBERT	.897	.815	.934	.590	.809
BERT-base	.885	.805	.927	.522	.785

Table 5.7: Comparison of domain specific pre-trained language models trained by category. Trained by fine-tuning, fixing the following hyperpameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Input data is title+abstract. Results are reported as macro-average between 10-folds.

The four categories in BATRACIO constitute a set of research phases, where some are adjacent between them. As observed during dataset annotation, disagreement between expert annotator is concentrated between neighbouring categories, as Table 3.5 shows. In the diagram shown in Figure 5.2.5, we show confusion matrix by categories of the predicted values by fine-tuned PubMedBERT, averaged for the 10 folds. As we can see in the diagram, the automatic system also has to deal with the difficulty of differentiating between adjacent categories. The boundaries between these pairs of categories seem to be more confusing, because there is the higher rate of error.

Dealing with an imbalanced dataset and with neighbouring categories

As shown in the previous subsections, two of the main challenges that systems have to deal with BATRACIO are the imbalanced distribution of samples per category and a configuration of categories where some are semantically adjacent to others. Categories are not mutually independent, they

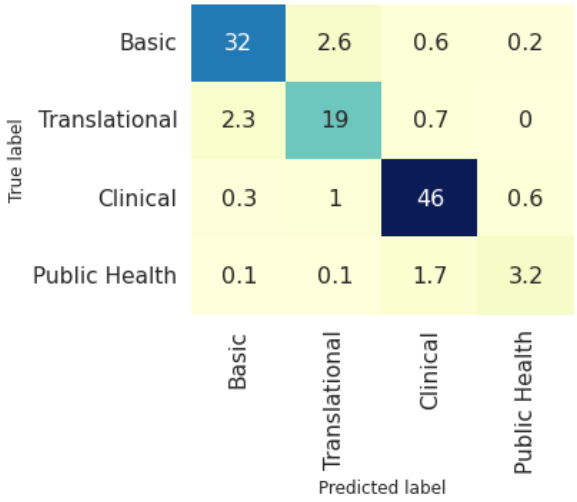


Figure 5.2: Averaged confusion matrix for fine-tuned version of PubMedBERT. Results are reported as average between 10-folds.

shape a semantic *value chain* and have semantic relations of neighbouring between them. Systems and domain experts have more difficulty discerning between adjacent categories, such as between `basic research` and `translational research`, because there is more ambiguity. In this section, as Table 5.8 shows, we explore whether task adaptation can be improved by addressing these two challenges during fine-tuning by modifying the loss function and introducing information about class distribution and adjacency between pairs of categories.

To address the category imbalance, we introduce a vector of weights, giving as weight to each category the inverse proportion of its frequency in the dataset, and to the loss function used during training. In this way, the inverse weight vector can should help to rescale the weight given to each class and would focus more on categories with fewer samples. For three of the models, it improves the F1 for the Public Health category and the overall F1 too. This approach could be a good option to mitigate the effect of the imbalance for the minority categories.

Another main challenge we find in this task is the fact of dealing with a configuration of categories where they form a semantic sequence relationship. Although it was not the aim of the project to explore this, and we have not found proposals for this specific problem in the literature. We have decided to probe the effect of reinforcing adjacency by modifying the loss function, as

	Basic	Trans.	Clinical	Pub. Health	ALL
System	**F1**	**F1**	**F1**	**F1**	**F1**
		Fine-tuning			
PubMedBERT	**.912**	**.850**	**.949**	**.688**	**.850**
sciBERT	.892	.816	.937	.639	.821
BioBERT	**.897**	**.815**	.934	.590	.809
BERT-base	.885	.805	.927	.522	.785
		+ Weighted loss			
PubMedBERT	.897	.826	.942	.659	.831
sciBERT	**.893**	**.819**	.938	.659	**.827**
BioBERT	.889	.796	.930	.628	.811
BERT-base	.879	.802	**.931**	**.609**	.805
		+ Neighbouring loss			
PubMedBERT	.904	.837	.942	.619	.826
sciBERT	.887	.813	**.939**	.621	.815
BioBERT	.893	.795	.932	.609	.807
BERT-base	.887	.799	**.931**	.606	**.806**
		+ Weighted and	*Neighbouring loss*		
PubMedBERT	.895	.828	.942	.635	.825
sciBERT	.891	.815	.938	**.663**	**.827**
BioBERT	.893	.805	**.938**	**.662**	**.825**
BERT-base	.885	.801	.930	.602	.804

Table 5.8: Comparison of domain specific pre-trained language models trained with modified loss function. Trained by fine-tuning, fixing the following hyperparameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Input data is title+abstract. Results are reported as macro-average between 10-folds.

neighbouring loss. Following the strategy of weighing categories with fewer samples in the cost function for imbalanced datasets (Tayyar Madabushi et al., 2019), we modify the loss function to consider the error to be double when categories are adjacent, thus it should reinforce adjacent categories to be separated during fine-tuning, given that as shown in Figure 5.2.5, this is where the highest error rate is concentrated. This approach performs well for BERT-base, but obtains similar or worse results for the other models.

Combining *weighted loss* and *neighbouring loss* seems to improve on the other configurations for sciBERT and BioBERT, to mitigate the two challenges of the dataset. However, the best PubMedBERT system does not improve, being the best configuration without modifying the loss function.

Balancing instances per category and removing "Public Health"

One of the particularities of our system is that it is imbalanced. In this section, we explore what would happen if our dataset was balanced or if we

did not have the minority class, "Public Health".

The results of these two experiments are available in the table 5.9.

	Removing "Public Health"		Balancing instances per category	
System	**F1**	**Acc.**	**F1**	**Acc.**
PubMedBERT	**.897**	**.914**	.756	.771
sciBERT	.892	.910	**.785**	**.791**
BioBERT	.889	.908	.624	.659
BERT-base	.880	.901	.649	.668

Table 5.9: Comparison of domain specific pre-trained language models for 3-category configuration and balanced-category configuration. Trained by fine-tuning, fixing the following hyperparameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Input data is title+abstract. Results are reported as macro-average between 10-folds.

If we remove the category "Public Health", the performance of the systems improves, because there are more samples and the three remaining categories are much more clear and robust. For instance, for the best system, PubMedBERT, the F-measure improves by 4.7 percentage points, and even more for the other three systems, bringing them closer to the performance of the best system.

However, if we balance categories to the minimum common number, in this case with 45 publications for training as "Public Health", for each category, the effectiveness per category is more comparable between categories, but the performance is much lower, due to the lower number of samples per category. Moreover, the number of samples per category is not representative enough to perform well in real applications.

5.2.6 Other experiments

In this last section, we explore if text pre-processing, as suggested in the literature (Schick and Schütze, 2019; Cenikj et al., 2021; Pascual et al., 2021; Ujiie et al., 2021), improves the capacity of biomedical pre-trained language models, as well as how hyperparameter setting contributes during training and how specific configurations can affect the learning capacity of models. For this, we explore the hyperparameters recommended in (Devlin et al., 2019).

Text cleaning

As identified in the literature, acronym resolution, removal of numbers and special characters, can improve performance of pre-trained language models in the biomedical domain. Table 5.10 compares the effect of cleaning abstracts, resolving acronyms, and removing numbers and special characters.

	Raw abstract		**Acronym**		**Num.+SC**		**Acr.+Num.+SC**	
System	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**
PubMedBERT	.813	.888	.813	.888	**.823**	**.892**	.812	.886
sciBERT	.804	.880	.805	.879	**.809**	**.882**	.804	.873
BioBERT	.794	.876	.811	.881	.801	.879	**.817**	**.882**
BERT-base	.772	.869	.784	.872	.766	.872	**.779**	**.875**

Table 5.10: Comparison of domain specific pre-trained language models with text cleaning. Trained by fine-tuning, fixing the following hyperameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Input data is title+abstract. Results are reported as macro-average between 10-folds. Column description: Acronym = resolving acronyms / Num.+SC = removing numbers and special characters / Acr.+Num.+SC = resolving acronyms, and removing numbers and special characters

Applying different pre-processing techniques helps systems to obtain more information from categories. However, not all preprocessing steps improve for all systems. For systems trained on BERT-base and BioBERT, which share vocabulary, acronym resolution improves, possibly because they do not have a vocabulary in the domain. However, for PubMedBERT and SciBERT, acronym resolution does not improve, because they already incorporate information in the domain that could include acronyms in many cases. The removal of numbers and special characters, which do not provide domain information and add basically noise, improves substantially for both models.

In conclusion, texts cleaning to resolve acronyms and remove numbers and special characters (frequent in scientific abstracts as results or formulas) can improve how pre-trained models process texts in the biomedical domain. However, it will have to be studied in detail how text cleaning affects the performance of our best configurations.

Exploring optimal hyperparameters

A major challenge in deep-learning based approaches is the selection of the best hyperparameters. Doing an exhaustive exploration of hyperparameter combination can be especially costly. Another big challenge, especially with small datasets, is overfitting. In other words, when the system learns exactly the training set, but does not learn to generalise the categories. This is very frequent and common due to the large number of parameters trained. In this section, we explore what the best hyperparameters are for the PubMedBERT model. The graphic in Figure 5.3 shows the loss rate of the training and validation sets during training for each epoch. The original BERT paper recommends a number of epochs between 2 and 4, as we can see in the graph, after 4 epochs the system does not learn and starts overfitting, getting worse when it evaluates with data it has not trained on. Here, we corroborate how the recommended number of epochs is also adequate for our dataset.

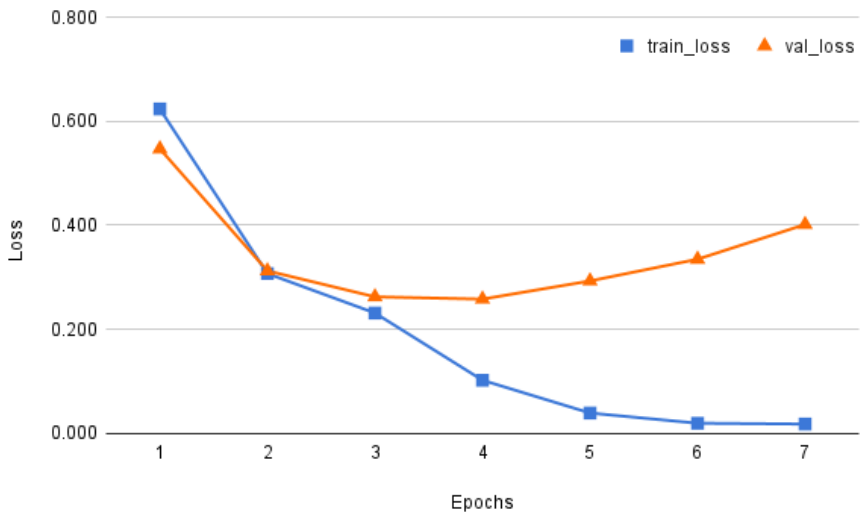


Figure 5.3: Averaged training and validation loss for 10-folds, with PubMedBERT.

During these learning epochs, the test set, which is not involved during training, has also been evaluated at each checkpoint, as shown in Figure 5.2.6. Indeed, the system improves in F1 and accuracy up to 4 epochs. After the fourth training epoch, it does not learn any more, and only learns

the training set. As observed in the literature, the use of an early stopping strategy can be a good solution to prevent overfitting.

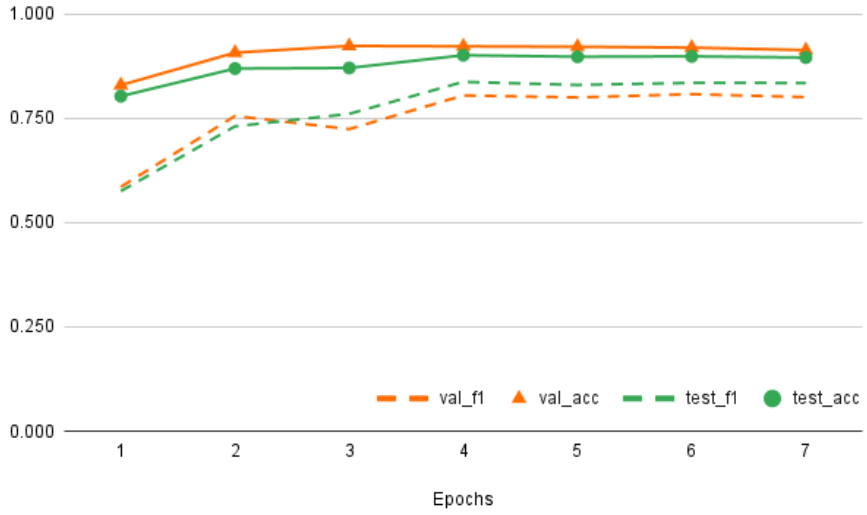


Figure 5.4: Averaged F1 and accuracy per epoch for 10-folds, with PubMedBERT.

Another fundamental hyperparameter is the learning rate. This is a tuning parameter in the optimisation algorithm that determines the step size of the movement towards a minimum in the loss function. If it is too large, the system will not be able to learn properly, but if it is too small, it may not learn enough. Table 5.11 explores the three learning rates recommended by (Devlin et al., 2019). With a lower learning rate, the models seem to learn better, however for our best model, PubMedBERT, we obtain the best results with a learning rate of 2e-5, the most used in the literature. For the other models, the improvement is greater with lower learning rates. Probably due to the complexity of the task, a lower learning rate can help to identify the categories more concisely.

	lr=2e-5		**lr=3e-5**		**lr=5e-5**	
System	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**
PubMedBERT	**.850**	**.907**	.835	.898	.844	.897
sciBERT	.821	.887	.826	.893	**.834**	**.894**
BioBERT	.809	.888	**.828**	**.895**	.822	.886
BERT-base	.785	0.878	.809	.882	**.826**	**.883**

Table 5.11: Comparison of learning rates in domain specific pre-trained language models. Trained by fine-tuning, fixing the following hyperameter configuration: 4 epochs and batch size of 16. Input data is title+abstract. Results are reported as macro-average between 10-folds.

Chapter 6

Conclusions and future work

This chapter summarises the main conclusions of the work carried out, and proposes some lines of future work.

6.1 Conclusions

The main objective of this project was to explore the automatic text classification of scientific publications according to the different phases in biomedical research, and to explore the use of pre-trained language models. The principal contribution of this work is the task and dataset of BATRACIO (*BAsic-TRAnslational-Clinical research phases classification in bIOmedical publications*). To the best of our knowledge, no previous work before has addressed the problem of automatically classifying scientific literature according to biomedical research phases.

The task seeks to help policymakers or funding agencies to better understand what research activities were carried out, mapping stakeholders and their research competencies, and hence to better allocate the resources, by classifying the scientific outputs of specific funding instruments or scientific publications, according to research phase. This information can also be used to better steer research funds towards proposals that are more appropriate to the ecosystem or which would probably have greater impact, avoiding the creation of funding gaps in different research phases and reducing the risk of duplicating efforts. Furthermore, it can also benefit universities in understanding which areas they are active in practice or to better priority-setting for their research strategy; or for researchers, to most easily extract information from large collections of scientific articles. As a result, the dataset

and systems can be used as a basis for exploration in real applications.

Since the problem proposed is new, the work described in this research includes the creation and annotation of a dataset of 1,100 scientific publications in the biomedical domain extracted from PubMed, categorized by research phase with the following categories: `basic research`, `translational research`, `clinical research`, and `public health`. Designing a new task is a big challenge, especially in a complex domain as biomedicine. For this reason, we have involved domain experts for designing the task and for annotating the dataset. However, the creation of the dataset has been costly and has required the organisation of several workshops for discussion with experts, data extraction and analysis in depth to reduce biases and to get publications in the domain of interest.

We have also explored whether the problem can be addressed automatically with general-domain systems, or if recent pre-trained language models improve on traditional methods. We have seen that state-of-the-art BERT-based pre-trained language models offer a great opportunity to solve this task. Different language models pre-trained from the open domain to the specific domain have been explored, and the results have been compared to other machine-learning techniques to assess whether the powerful tool of BERT, and specifically with pre-trained models in the biomedical or scientific domain, obtains better results and whether a better generalisation of the context improves the solving of the task proposed. Furthermore, we have also explored how to use them for text classification and which strategies may be favourable for the classification of biomedical research articles, such as text cleaning and hyperparameter setting.

Different domain-specific models in the biomedical domain have appeared in recent years, trying to incorporate domain knowledge through different approaches, although not all models work equally well for this task. Furthermore, as seen in the experiments, minor modifications in the configuration of hyperparameters, in text cleaning[1] or in the input data, can change the performance of the systems.

We have seen that using pre-trained language models and adapting them to a downstream task of text classification by fine-tuning, adding a classification layer on top of the model and updating parameter weights, allows for a reduction in the resource consumption, because training a model from

[1]Such as acronym resolution, or number and special character elimination.

scratch requires a lot of computational resources and execution time. The community-driven paradigm and thanks to the HuggingFace community, it is more feasible and easier to use these Transformer-based systems.

Nevertheless, the main specific challenges of our dataset are the class imbalance and that categories are not mutually independent, they shape a semantic *value chain* and have semantic relations of adjacency between them. This was not a main goal of the project, but we have also explored whether slight modifications in the loss function can deal with imbalanced and adjacent categories. Although the results of these experiments are partially satisfactory, they point to future lines of research.

6.2 Future work

This work has opened up new lines of research that could be tackled in the future.

As future lines of work, it could be interesting to look with domain experts at the particular cases where systems fail and to try to understand the source of those errors, if there is some kind of bias, or documents are more complicated or diffuse. In addition, in the next months, the best system will be applied to two real applications, two collections of scientific publications. One collection is in the domain of cancer research, which introduces the challenge of portability of the system to a much more specific domain. The other collection concentrates on publications of a university hospital. The use of the system in these real applications will be evaluated by experts in order to understand its real effectiveness outside the experimental and evaluation environment, limited by the dataset.

The dataset developed is small, and a possible line of improvement could be to explore data augmentation techniques to increase samples of categories with fewer samples. However, as the experiment results point out, the best solution would be to increase the number of samples in the dataset, in order to increase the absolute number of samples of minority categories.

In this project we have explored as data inputs only the title and abstract, the minimum unit of annotation, according to experts. However, there is an additional interesting research line related to exploiting other sections of scientific articles, such as *Introduction* or *Materials and Methods* section, where there may be relevant information about the research phase.

Another interesting direction is to incorporate the semantic relation between categories using other unexplored techniques. In most cases, categories or classes are symbolic labels and no additional information about them is available, such as introducing representation in the pre-trained model of the category, other kind of external information, or even exploring different modifications in the loss function of the performed in this project.

In the creation of the dataset, publications in the biomedical domain have been filtered by a rule-based classification approach with ontological information from MeSH taxonomy. The creation of a dataset for identifying which publication belong to the biomedical domain could allow the application of the designed system given any publication and not only in the biomedical domain. Another challenge identified during the creation of the dataset, in which only scientific articles have been filtered, is that some review articles are reported as articles, but they are in fact reviews of techniques and do not involve a research activity to be analysed, but rather the review of techniques or research on diseases across different research phases. One possible approach to differentiate between articles and reviews is to create a trained system with articles and reviews, which would not require the contribution of experts.

As can be seen in the literature, increasing the model size lead to improve performance of tasks. In recent years, larger models, with more parameters, more robust, trained on larger collections, are continuously appearing. However, there are not many adapted to the biomedical domain. Another line of exploration is to use systems on pre-trained language models with larger architectures than BERT.

We have adapted pre-trained language models with fine-tuning, but recent proposals include other downstream task adaptation techniques, such as p-tunning or prompt-tunning, which offer more efficient and faster ways to adapt the models. However, they also present many challenges that should be addressed.

Finally, immediate future work will include writing a scientific journal article to communicate and share the results of this work with the research community.

Bibliography

Bibliography

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2016:310–318.

Asan Agibetov, Kathrin Blagec, Hong Xu, and Matthias Samwald. 2018. Fast and scalable neural embedding models for biomedical sentence classification. *BMC Bioinformatics*, 19:541.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Mohammad Anani, Nazmul Kazi, Matthew Kuntz, and Indika Kahanda. 2019. RDoC task at BioNLP-OST 2019. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 216–226, Hong Kong, China. Association for Computational Linguistics.

Éric Archambault, Olivier H. Beauchesne, and J. Caruso. 2013. Towards a multilingual, comprehensive and open scientific journal ontology.

Yusuf Arslan, Kevin Allix, Lisa Veiber, Cedric Lothritz, Tegawendé F. Bissyandé, Jacques Klein, and Anne Goujon. 2021. A comparison of pre-trained language models for multi-class text classification in the financial domain. *Companion Proceedings of the Web Conference 2021*.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather L. Butler, J. Michael Cherry, Allan Peter Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna E. Lewis, John C. Matese,

Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Asmaa Aubaid and Alok Mishra. 2020. A rule-based approach to embedding techniques for text document classification. *Applied Sciences*, 10.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32 3:432–40.

Adrian Barnett and Zoe Doubleday. 2020. Meta-research: The growth of acronyms in the scientific literature. *eLife*, 9:e60080.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. Cometa: A corpus for medical entity linking in the social media.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. Association for Computational Linguistics.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of*

the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 238–251, New Orleans, Louisiana. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Karthiek Reddy Bokka, Shubhangi Hora, Tanuj Jain, and Monicah Wambugu. 2019. *Deep Learning for Natural Language Processing: Solve your natural language processing problems with smart deep neural networks*. Packt Publishing Ltd.

Gully Burns, Xiangci Li, and Nanyun Peng. 2019. Building deep learning models for evidence classification from the open access biomedical literature. *Database: The Journal of Biological Databases and Curation*, 2019.

D. Butler. 2008. Translational research: crossing the valley of death. In *Nature*.

Kathi Canese and Sarah Weis. 2013. Pubmed: The bibliographic database.

William Cavnar and John Trenkle. 2001. N-gram-based text categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*.

Gjorgjina Cenikj, Tome Eftimov, and Barbara Koroušić Seljak. 2021. SAFFRON: tranSfer leArning for food-disease RelatiOn extractioN. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 30–40, Online. Association for Computational Linguistics.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679,

Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wendy Chapman and Kevin Cohen. 2009. Current issues in biomedical text mining and natural language processing. *Journal of biomedical informatics*, 42:757–9.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers.

Kevin Cohen. 2014. *Biomedical Natural Language Processing and Text Mining*, pages 141–177.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch.

Wikimedia Commons. 2020. File:british reptiles and batrachians (1888) (14781336724).jpg — wikimedia commons, the free media repository. [Online; accessed 30-March-2022].

S. Crawford-Welch and K. McCleary. 1992. An identification of the subject areas and research techniques used in five hospitality-related journals. *International Journal of Hospitality Management*, 11:155–167.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2018. *Proceedings of the BioNLP 2018 workshop*. Association for Computational Linguistics, Melbourne, Australia.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2021. *Proceedings of the 20th Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Andres Duque, R.M. Stevenson, Juan Martinez-Romo, and Lourdes Araujo. 2018. Co-occurrence graphs for word sense disambiguation in the biomedical domain. *Artificial Intelligence in Medicine*, 87.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. SemEval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, Denver, Colorado. Association for Computational Linguistics.

Holly Else. 2020. How a torrent of covid science changed research publishing — in seven charts. *Nature*, 588:553–553.

Hermenegildo Fabregat, Juan Martinez-Romo, and Lourdes Araujo. 2020. Understanding and improving disability identification in medical documents. *IEEE Access*, PP:1–1.

Jeffrey Flier and Joseph Loscalzo. 2017. Categorizing biomedical research: The basics of translation. *The FASEB Journal*, 31:3210–3215.

Daniel Fort, Timothy Herr, Pamela Shaw, Karen Gutzman, and Justin Starren. 2017. Mapping the evolving definitions of translational research. *Journal of Clinical and Translational Science*, 1:1–7.

Carol Friedman, Pauline Kra, and A. Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of biomedical informatics*, 35 4:222–35.

Enric Fuster, Francesco Massucci, and Monika Matusiak. 2020. Identifying specialisation domains beyond taxonomies: mapping scientific and technological domains of specialisation via semantic analyses. In Roberta Capello, Alexander Kleibrink, and Monika Matusiak, editors, *Quantitative Methods for Place-Based Innovation Policy*, page 195–234.

John Michael Giorgi and Gary D Bader. 2019. Towards reliable named entity recognition in the biomedical domain. *bioRxiv*.

Lorraine Goeuriot, Hanna Suominen, Liadh Kelly, Laura Alemany, Nicola Brew-Sam, Viviana Cotik, Dario Filippo, Gabriela Gonzalez-Sáez, Franco Luque, Philippe Mulhem, Gabriella Pasi, Roland Roller, Sandaru Seneviratne, Jorge Vivaldi, Marco Viviani, and Chenchen Xu. 2021a. *CLEF eHealth Evaluation Lab 2021*, pages 593–600.

Lorraine Goeuriot, Hanna Suominen, Gabriella Pasi, Elias Bassani, Nicola Brew-Sam, Gabriela Nicole González Sáez, Liadh Kelly, Philippe Mulhem,

Sandaru Seneviratne, Rishabh Upadhyay, Marco Viviani, and Chenchen Xu. 2021b. Consumer health search at clef ehealth 2021. In *CLEF*.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Zhengyi Guan and Renyuan Liu. 2021. Yunnan-deep at ehealth-kd challenge 2021: Deep learning model for entity recognition in spanish documents. In *IberLEF@SEPLN*.

Stephen Hanney, Sophie Castle-Clarke, Jonathan Grant, Susan Guthrie, Chris Henshall, Jorge Mestre-Ferrandiz, Michele Pistollato, Alexandra Pollitt, Jon Sussex, and Steven Wooding. 2015. How long does biomedical research take? studying the time taken between biomedical and health research and its translation into products, policy, and practice. *Health research policy and systems / BioMed Central*, 13:1.

Rave Harpaz, Alison Callahan, Suzanne R. Tamang, Yen S. Low, David J. Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam Haresh Shah. 2014. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Safety*, 37:777–790.

Zellig Sabbettai Harris. 1991. Theory of language and information: a mathematical approach.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. pages 4604–4614.

Michael Head, Joseph Fitchett, Rifat Atun, and Robin May. 2014. Systematic analysis of funding awarded for mycology research to institutions in the uk, 1997–2010. *BMJ open*, 4:e004129.

William Hogan, Yoshiki Vazquez Baeza, Yannis Katsis, Tyler Baldwin, Ho-Cheol Kim, and Chun-Nan Hsu. 2021. BLAR: Biomedical local acronym resolver. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 126–130, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification.

Chung-Chi Huang and Zhiyong Lu. 2015. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in Bioinformatics*, 17(1):132–144.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Thorsten Joachims. 1998. Text categorization with support vector machines. *Proc. European Conf. Machine Learning (ECML'98)*.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics*, 126:103982.

V. Kieuvongngam, Bowen Tan, and Yiming Niu. 2020. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. *ArXiv*, abs/2006.01997.

Seongsoon Kim, Donghyeon Park, Yonghwa Choi, Kyubum Lee, Byounggun Kim, Minji Jeon, Jihye Kim, Aik Choon Tan, and Jaewoo Kang. 2018. A pilot study of biomedical text comprehension using an attention-based deep neural reader: Design and experimental analysis. *JMIR Medical Informatics*, 6.

Youngjoong Ko and Jungyun Seo. 2000. Automatic text categorization by unsupervised learning. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, page 453–459, USA. Association for Computational Linguistics.

Dimitrios Koutsomitropoulos and Andreas Andriopoulos. 2020. *Automated MeSH Indexing of Biomedical Literature Using Contextualized Word Representations*, pages 343–354.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, Donald Brown, Laura Id, and Barnes. 2019. Text classification algorithms: A survey. *Information (Switzerland)*, 10.

Lissy Laar, Thijs Kruif, Ludo Waltman, Ingeborg Meijer, Anshu Gupta, and Niels Hagenaars. 2018. Improving the evaluation of worldwide biomedical research output: Classification method and standardised bibliometric indicators by disease. *BMJ Open*, 8:e020818.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org.

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *J. of Biomedical Informatics*, 57(C):28–37.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Changchun Li, Ximing Li, and Jihong Ouyang. 2021. Semi-supervised text classification with balanced deep representation distributions. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5044–5053, Online. Association for Computational Linguistics.

Dingcheng Li, Sijia Liu, Majid Rastegar-Mojarad, Yanshan Wang, Vipin Chaudhary, Terry Therneau, and Hongfang Liu. 2017. A topic-modeling based framework for drug-drug interaction classification from biomedical text. *AMIA Annual Symposium Proceedings*, 2016:789–798.

Fei Li, Yonghao Jin, Weisong Liu, Bhanu Rawat, Pengshan Cai, and Hong Yu. 2019a. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: An empirical study. *JMIR Medical Informatics*, 7:e14830.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).

Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2019b. Behrt: Transformer for electronic health records.

Salvador Lima-López, Eulàlia Farré-Maduell, Antonio Miranda-Escalada, Vicent Briva-Iglesias, and Martin Krallinger. 2021. Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. *Proces. del Leng. Natural*, 67:243–256.

Carolyn E. Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88 3:265–6.

Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020. Comparative analysis of text classification approaches in electronic health records. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 86–94, Online. Association for Computational Linguistics.

Mariana Mazzucato. 2021. *MISSION ECONOMY: a moonshot guide to changing capitalism*. PENGUIN BOOKS.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

I. C. Mcllwaine and N. Williamson. 1999. International trends in subject analysis research. *Knowledge Organization*, 26:23–29.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning based text classification: A comprehensive review.

James G. Mork, Antonio Jimeno-Yepes, and Alan R. Aronson. 2013. The nlm medical text indexer system for indexing biomedical literature. In *BioASQ@CLEF*.

Ghulam Mujtaba, Liyana Shuib, Norisma Idris, Wai Lam Hoo, Ram Gopal Raj, Kamran Khowaja, Khairunisa Shaikh, and Henry Friday Nweke. 2019. Clinical text classification research trends: Systematic literature review and open issues. *Expert Systems with Applications*, 116:494–520.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Usman Naseem, Adam G. Dunn, Matloob Khushi, and Jinman Kim. 2021. Benchmarking for biomedical natural language processing tasks with a domain specific albert.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2021. Overview of BioASQ 2021: The ninth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In *Lecture Notes in Computer Science*, pages 239–263. Springer International Publishing.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models.

Karen Oconnor, Abeed Sarker, Jeanmarie Perrone, and Graciela Hernandez. 2020. Promoting reproducible research for characterizing nonmedical use of medications through data annotation: Description of a twitter corpus and guidelines. *Journal of medical Internet research*, 22:e15861.

Damian Pascual, Sandro Luck, and Roger Wattenhofer. 2021. Towards BERT-based automatic ICD coding: Limitations and opportunities. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 54–63, Online. Association for Computational Linguistics.

Pinal Patel, Disha Davey, and Parth Pathak. 2018. Annotation of a large clinical entity corpus. pages 2033–2042.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. volume 14, pages 1532–1543.

Bethany Percha. 2021. Modern clinical text mining: A guide and review.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks.

M. Porter. 1985. Competitive advantage: Creating and sustaining superior performance.

Sunil Prabhakar and Dong-Ok Won. 2021. Medical text classification using hybrid deep learning models with multihead attention. *Computational Intelligence and Neuroscience*, 2021:1–16.

Li Qing, Weng Linhong, and Ding Xuehai. 2019. A novel neural network-based method for medical text classification. *Future Internet*, 11:255.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and Xuan-Jing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Philip Resnik, Katherine E. Goodman, and Mike Moran. 2020. Developing a curated topic model for COVID-19 medical research literature.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Nayat Sanchez-Pi, Luis Martí, and Ana Cristina Bicharra Garcia. 2016. Improving ontology-based text classification: An occupational health and security application. *Journal of Applied Logic*, 17:48–58. SOCO13.

Mourad Sarrouti and Said Ouatik El Alaoui. 2020. Sembionlqa: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artificial intelligence in medicine*, 102:101767.

Eric W Sayers, Richa Agarwala, Evan E Bolton, J Rodney Brister, Kathi Canese, Karen Clark, Ryan Connor, Nicolas Fiorini, Kathryn Funk, Timothy Hefferon, J Bradley Holmes, Sunghwan Kim, Avi Kimchi, Paul A Kitts, Stacy Lathrop, Zhiyong Lu, Thomas L Madden, Aron Marchler-Bauer, Lon Phan, Valerie A Schneider, Conrad L Schoch, Kim D Pruitt, and James Ostell. 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 47(D1):D23–D28.

Timo Schick and Hinrich Schütze. 2019. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking.

Ariel Schwartz and Marti Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 4:451–62.

Fabrizio Sebastiani. 2001. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Zhengfei Shen and Shaohua Zhang. 2020. A novel deep-learning-based model for medical text classification. pages 267–273.

Victoria Simpkin, Matthew Renwick, Ruth Kelly, and Elias Mossialos. 2017. Incentivising innovation in antibiotic drug discovery and development: Progress, challenges and next steps. *The Journal of Antibiotics*, 70.

Peng Su, Yifan Peng, and K. Vijay-Shanker. 2021. Improving bert model using contrastive learning for biomedical relation extraction. In *BIONLP*.

Yuhan Su, Hongxin Xiang, Haotian Xie, Yong Yu, Shiyan Dong, Zhaogang Yang, and Na Zhao. 2020. Application of bert to enable gene classification based on clinical evidence. *BioMed Research International*, 2020:1–13.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *CCL*.

Zijun Sun, Chun Fan, Xiaofei Sun, Yuxian Meng, Fei Wu, and Jiwei Li. 2020. Neural semi-supervised learning for text classification under large-scale pretraining. *arXiv preprint arXiv:2011.08626*.

Madhumita Sushil, Simon Suster, and Walter Daelemans. 2021. Are we there yet? exploring clinical domain knowledge of BERT models. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 41–53, Online. Association for Computational Linguistics.

György Szarvas, Veronika Vincze, Richárd Farkas, and Janos Csirik. 2008. The bioscope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. *BioNLP*, pages 38–45.

Madis Talmar, Bob Walrave, Ksenia Podoynitsyna, Jan Holmström, and Georges Romme. 2020. Mapping, analyzing and designing innovation ecosystems: The ecosystem pie model. *Long Range Planning*, 53:101850.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China. Association for Computational Linguistics.

Sai Krishna Telukuntla, Aditya Kapri, and Wlodek Zadrozny. 2020. Uncc biomedical semantic question answering systems. bioasq: Task-7b, phase-b.

Jalaj Thanaki. 2017. *Python natural language processing*. Packt Publishing Ltd.

Emanuela Todeva and Ruslan Rakhmatullin. 2016. Global value chains mapping: Methodology and cases for policy makers.

Vu D. Tran, Van-Hien Tran, Phuong Minh Nguyen, C. Nguyen, K. Satoh, Yuji Matsumoto, and Minh Nguyen. 2021. Covrelex: A covid-19 retrieval system with relation extraction.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers*. " O'Reilly Media, Inc.".

Shogo Ujiie, Hayate Iso, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2021. End-to-end biomedical entity linking with span-based dictionary matching. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 162–167, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2021. Pre-training technique to localize medical bert and enhance biomedical bert.

Nancy Wang, Diwakar Mahajan, and Sara Rosenthal. 2021. Semeval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (sem-tab-facts).

Yuxuan Wang, Yutai Hou, Wanxiang Che, and Ting Liu. 2020. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, 11.

Griffin Weber. 2013. Identifying translational science within the triangle of biomedicine. *Journal of translational medicine*, 11:126.

Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Hua Xu. 2020. Relation extraction from clinical narratives using pre-trained language models. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2019:1236–1245.

S. Woolf. 2008. The meaning of translational research and why it matters. *JAMA*, 299 2:211–3.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu. 2019. Deep learning in clinical natural language processing: A methodical review. *Journal of the American Medical Informatics Association*, 27.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2018a. Clinical text classification with rule-based features and knowledge-guided convolutional neural

networks. *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, pages 70–71.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2018b. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links.

Antonio Jose Jimeno Yepes, Laura Plaza, Jorge Carrillo-de Albornoz, James G Mork, and Alan R Aronson. 2015. Feature engineering for medline citation categorization with mesh. *BMC bioinformatics*, 16(1):1–12.

Zhang Yijia, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019a. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019b. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

Ronghui You, Yuxuan Liu, Hiroshi Mamitsuka, and Shanfeng Zhu. 2020. Bertmesh: Deep contextual representation learning for large-scale high-performance mesh indexing with full text. *Bioinformatics (Oxford, England)*.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190, Online. Association for Computational Linguistics.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1:43–52.

Minghao Zhu and Keyuan Jiang. 2021. Semi-supervised language models for identification of personal health experiential from Twitter data: A case for medication effects. In *Proceedings of the 20th Workshop on Biomedical*

Language Processing, pages 228–237, Online. Association for Computational Linguistics.

Appendix A

MeSH Filter defining the scope of biomedicine

The next table contains in a CSV format the taxonomical filters used to retrieve publications in the domain of biomedical research from their MeSH terms.

```
Scope,Main MeSH Heading,Tree Number / ID,Add,Exclude MESH,Exclude TREE
Biomedical Research,Anatomy ,A,TRUE,,
Biomedical Research,Diseases ,C,TRUE,,
Biomedical Research,Chemicals and Drugs,D,TRUE,,
Biomedical Research,"Analytical, Diagnostic and Therapeutic Techniques, and Equipmen
Biomedical Research,Health Care ,N,,,TRUE
Biomedical Research,Bioengineering ,J01.293.069,TRUE,,
Biomedical Research,Biomedical Engineering ,J01.293.140,TRUE,,
Biomedical Research,Health Occupations ,H02,,,TRUE
Biomedical Research,Statistics as Topic ,E05.318,,,TRUE
Biomedical Research,"Models, Theoretical ",E05.599,,,TRUE
Biomedical Research,"Imaging, Three-Dimensional",E01.370.350.400,,,TRUE
Biomedical Research,Equipment Design ,D004867,,TRUE,
Biomedical Research,Equipment Failure Analysis,D019544,,TRUE,
Biomedical Research,Finite Element Analysis,D020342,,TRUE,
Biomedical Research,Equipment Safety,D004869,,TRUE,
Biomedical Research,Fourier Analysis,D005583,,TRUE,
Biomedical Research,Rheology ,E05.830,,,TRUE
Biomedical Research,"Games, Experimental ",E05.385,,,TRUE
Biomedical Research,Thermometry,E05.933,,,TRUE
Biomedical Research,Electrical Equipment and Supplies,E07.305,,,TRUE
Biomedical Research,Brain-Computer Interfaces ,E07.305.076,TRUE,,
```

```
Biomedical Research,Optical Devices ,E07.632,,,TRUE
Biomedical Research,"Phantoms, Imaging",D019047,,TRUE,
Biomedical Research,Quantum Dots,D045663,,TRUE,
Biomedical Research,Radiation Equipment and Supplies ,E07.710,,,TRUE
Biomedical Research,Bioengineering,D057005,,TRUE,
Biomedical Research,Synthetic Biology,D058615,,TRUE,
Biomedical Research,Electrodes,D004566,,TRUE,
Biomedical Research,Biota ,N06.230.124.049.100,TRUE,,
Biomedical Research,Disease Vectors,N06.850.310.350,TRUE,,
Biomedical Research,Drug Contamination,N06.850.360,TRUE,,
Biomedical Research,Molecular Epidemiology,E05.318.416,TRUE,,
Biomedical Research,Organic Chemicals,D02,,,TRUE
Biomedical Research,Proteins,D011506,,TRUE,
Biomedical Research,Specialty Uses of Chemicals,D27.720,,,TRUE
Biomedical Research,Extremities,A01.378,,,TRUE
Biomedical Research,Head,A01.456,,,TRUE
Biomedical Research,Inorganic Chemicals , D01,,,TRUE
Biomedical Research,Brain,D001921,,TRUE,
Biomedical Research,Polymers,D25.720,,,TRUE
Biomedical Research,Polymers,D05.750,,,TRUE
Biomedical Research,Oils,D10.627,,,TRUE
Biomedical Research,Complex Mixtures ,D20,,,TRUE
Biomedical Research,Physical Examination,E01.370.600,,,TRUE
Biomedical Research,Complementary Therapies ,E02.190,,,TRUE
Biomedical Research,Artifacts,E05.047,,,TRUE
Biomedical Research,Breeding ,E05.820.150,,,TRUE
Biomedical Research,Psychological Techniques ,E05.796,,,TRUE
Biomedical Research,Weights and Measures,E05.978,,,TRUE
Biomedical Research,Telemetry,E05.925,,,TRUE
Biomedical Research,Phenomena and Processes,G,TRUE,,
Biomedical Research,Physical Phenomena,G01,,,TRUE
Biomedical Research,Chemical Phenomena,G02,,,TRUE
Biomedical Research,Plant Physiological Phenomena,G15,,,TRUE
Biomedical Research,Mathematical Concepts,G17,,,TRUE
Biomedical Research,Biological Phenomena,G16,,,TRUE
Biomedical Research,Plant Structures ,A18,,,TRUE
Biomedical Research,Viruses,[B04],TRUE,,
Biomedical Research,Plasmodium ,[B01.043.075.380.611],TRUE,,
Biomedical Research,Host Microbial Interactions,[G16.527],TRUE,,
Biomedical Research,Regeneration ,[G16.762],TRUE,,
Biomedical Research,"Remission, Spontaneous",[G16.767],TRUE,,
```

```
Biomedical Research,Organelle Biogenesis,[G16.645],TRUE,,
Biomedical Research,Neurosciences,[H01.158.610],TRUE,,
Biomedical Research,Biogenic Monoamines,[D02.092.211.215],TRUE,,
Biomedical Research,"Models, Biological",[E05.599.395],TRUE,,
Biomedical Research,"Models, Animal",[E05.598],TRUE,,
Biomedical Research,Cognition,D003071,TRUE,,
Biomedical Research,Neurocognitive Disorders,[F03.615],TRUE,,
Biomedical Research,Vaccines,[D20.215.894],TRUE,,
Biomedical Research,Toxoplasma,[B01.043.075.189.250.750.800],TRUE,,
Biomedical Research,Nanotechnology,[H01.603],TRUE,,
Biomedical Research,Biomedical Technology,[J01.897.115],TRUE,,
Biomedical Research,Mycobacterium tuberculosis,D009169,TRUE,,
Biomedical Research,"Specialties, Surgical",[H02.403.810],TRUE,,
Biomedical Research,Aerosols,D000336,,TRUE,
Biomedical Research,Greenhouse Gases,D000074382,,TRUE,
Biomedical Research,Photosynthesis,D010788,,TRUE,
Biomedical Research,Plant Cells,[A11.750],,,TRUE
Biomedical Research,Movement,[G11.427.410],,,TRUE
Biomedical Research,"Yeast, Dried",D015002,,TRUE,
Biomedical Research,Photosystem II Protein Complex,D045332,,TRUE,
Biomedical Research,Photosystem I Protein Complex,D045331,,TRUE,
Biomedical Research,Light-Harvesting Protein Complexes,D045342,,TRUE,
Biomedical Research,Feeding Behavior,[G07.203.650.353],,,TRUE
Biomedical Research,Alloys,D000497,,TRUE,
Biomedical Research,Oxidation-Reduction,D010084,,TRUE,
Biomedical Research,Bioreactors,[E07.115],,,TRUE
Biomedical Research,Cellulose,D002482,,TRUE,
Biomedical Research,Cell Wall,D002473,,TRUE,
Biomedical Research,Phonation,G09.772.585,,,TRUE
Biomedical Research,Voice,G09.772.925,,,TRUE
Biomedical Research,Parturition,D036801,,TRUE,
Biomedical Research,Sign language,D012813,,TRUE,
Biomedical Research,Hearing,D006309,,TRUE,
Biomedical Research,Diet,D004032,,TRUE,
Biomedical Research,Beverages,G07.203.100,,,TRUE
Biomedical Research,Fermented Foods,G07.203.200,,,TRUE
Biomedical Research,Food,G07.203.300,,,TRUE
Biomedical Research,Breeding,D001947,,TRUE,
Biomedical Research,Selective Breeding,D000068618,,TRUE,
Biomedical Research,"DNA, Ancient",D000072441,,TRUE,
Biomedical Research,Phylogeny,D010802,,TRUE,
```

```
Biomedical Research,Biological Evolution ,G16.075,,,TRUE
Biomedical Research,Biological Evolution ,G05.045,,,TRUE
Biomedical Research,Gene Frequency ,G05.330,,,TRUE
Biomedical Research,Genetic Variation ,G05.365,,TRUE,
Biomedical Research,Radiometric Dating,D055110,,TRUE,
```

Appendix B

Subject domain filter based on the Science-Metrix Taxonomy

The next table contains in a CSV format the taxonomical filters used to retrieve publications in the domain of biomedical research from the subject domain classification of their journal.

```
Field_English,SubField_English,REMOVE
"Agriculture, Fisheries & Forestry",Veterinary Sciences,REMOVE
"Agriculture, Fisheries & Forestry",Dairy & Animal Science,REMOVE
"Agriculture, Fisheries & Forestry",Food Science,REMOVE
"Agriculture, Fisheries & Forestry",Fisheries,REMOVE
"Agriculture, Fisheries & Forestry",Agronomy & Agriculture,REMOVE
Biology,Plant Biology & Botany,REMOVE
Biology,Evolutionary Biology,REMOVE
Biology,Entomology,REMOVE
Biology,Zoology,REMOVE
Biology,Ecology,REMOVE
Biology,Marine Biology & Hydrobiology,REMOVE
Biomedical Research,Biochemistry & Molecular Biology,KEEP
Biomedical Research,Microbiology,KEEP
Biomedical Research,Developmental Biology,KEEP
Biomedical Research,Virology,KEEP
Biomedical Research,Toxicology,KEEP
Biomedical Research,Physiology,KEEP
Biomedical Research,Nutrition & Dietetics,KEEP
```

Biomedical Research, Genetics & Heredity, KEEP
Biomedical Research, Biophysics, KEEP
Biomedical Research, Mycology & Parasitology, KEEP
Biomedical Research, Anatomy & Morphology, KEEP
Biomedical Research, Microscopy, KEEP
Built Environment & Design, Building & Construction, REMOVE
Chemistry, Medicinal & Biomolecular Chemistry, KEEP
Chemistry, Organic Chemistry, KEEP
Chemistry, Analytical Chemistry, KEEP
Chemistry, General Chemistry, KEEP
Chemistry, Polymers, KEEP
Chemistry, Inorganic & Nuclear Chemistry, KEEP
Chemistry, Physical Chemistry, REMOVE
Clinical Medicine, Oncology & Carcinogenesis, KEEP
Clinical Medicine, Neurology & Neurosurgery, KEEP
Clinical Medicine, Cardiovascular System & Hematology, KEEP
Clinical Medicine, Immunology, KEEP
Clinical Medicine, General & Internal Medicine, KEEP
Clinical Medicine, Nuclear Medicine & Medical Imaging, KEEP
Clinical Medicine, Surgery, KEEP
Clinical Medicine, Obstetrics & Reproductive Medicine, KEEP
Clinical Medicine, Pharmacology & Pharmacy, KEEP
Clinical Medicine, Orthopedics, KEEP
Clinical Medicine, Gastroenterology & Hepatology, KEEP
Clinical Medicine, Endocrinology & Metabolism, KEEP
Clinical Medicine, Urology & Nephrology, KEEP
Clinical Medicine, Ophthalmology & Optometry, KEEP
Clinical Medicine, Dentistry, KEEP
Clinical Medicine, Dermatology & Venereal Diseases, KEEP
Clinical Medicine, Respiratory System, KEEP
Clinical Medicine, Pediatrics, KEEP
Clinical Medicine, Otorhinolaryngology, KEEP
Clinical Medicine, Psychiatry, KEEP
Clinical Medicine, Arthritis & Rheumatology, KEEP
Clinical Medicine, Anesthesiology, KEEP
Clinical Medicine, Pathology, KEEP
Clinical Medicine, Emergency & Critical Care Medicine, KEEP
Clinical Medicine, General Clinical Medicine, KEEP
Clinical Medicine, Sport Sciences, KEEP
Clinical Medicine, Tropical Medicine, KEEP
Clinical Medicine, Allergy, KEEP

Clinical Medicine,Geriatrics,KEEP
Clinical Medicine,Complementary & Alternative Medicine,KEEP
Clinical Medicine,Legal & Forensic Medicine,KEEP
Clinical Medicine,Environmental & Occupational Health,KEEP
Earth & Environmental Sciences,Environmental Sciences,KEEP
Earth & Environmental Sciences,Meteorology & Atmospheric Sciences,REMOVE
Earth & Environmental Sciences,Geochemistry & Geophysics,REMOVE
Economics & Business ,Logistics & Transportation,REMOVE
Economics & Business ,Business & Management,REMOVE
Enabling & Strategic Technologies,Biotechnology,KEEP
Enabling & Strategic Technologies,Nanoscience & Nanotechnology,KEEP
Enabling & Strategic Technologies,Bioinformatics,KEEP
Enabling & Strategic Technologies,"Strategic, Defence & Security Studies",REMOVE
Enabling & Strategic Technologies,Optoelectronics & Photonics,KEEP
Engineering,Biomedical Engineering,KEEP
Engineering,Environmental Engineering,REMOVE
Engineering,Operations Research,REMOVE
"General Arts, Humanities & Social Sciences","General Arts, Humanities & Social Scie
General Science & Technology,General Science & Technology,KEEP
Historical Studies,Anthropology,REMOVE
Historical Studies,"History of Science, Technology & Medicine",REMOVE
Information & Communication Technologies,Medical Informatics,KEEP
Information & Communication Technologies,Artificial Intelligence & Image Processing,
Mathematics & Statistics,Statistics & Probability,KEEP
Philosophy & Theology,Applied Ethics,KEEP
Physics & Astronomy,Chemical Physics,REMOVE
Physics & Astronomy,Acoustics,KEEP
Physics & Astronomy,Optics,KEEP
Physics & Astronomy,Fluids & Plasmas,KEEP
Physics & Astronomy,General Physics,KEEP
Physics & Astronomy,Astronomy & Astrophysics,KEEP
Physics & Astronomy,Applied Physics,KEEP
Psychology & Cognitive Sciences,Experimental Psychology,KEEP
Psychology & Cognitive Sciences,Developmental & Child Psychology,KEEP
Psychology & Cognitive Sciences,Clinical Psychology,KEEP
Psychology & Cognitive Sciences,Behavioral Science & Comparative Psychology,KEEP
Psychology & Cognitive Sciences,Social Psychology,KEEP
Psychology & Cognitive Sciences,Human Factors,KEEP
Psychology & Cognitive Sciences,General Psychology & Cognitive Sciences,KEEP
Psychology & Cognitive Sciences,Psychoanalysis,KEEP
Public Health & Health Services,Public Health,KEEP

Public Health & Health Services,Rehabilitation,KEEP
Public Health & Health Services,Nursing,KEEP
Public Health & Health Services,Epidemiology,KEEP
Public Health & Health Services,Health Policy & Services,KEEP
Public Health & Health Services,Gerontology,KEEP
Public Health & Health Services,Substance Abuse,KEEP
Public Health & Health Services,Speech–Language Pathology & Audiology,KEEP
Social Sciences,Education,KEEP
Social Sciences,Criminology,KEEP
Social Sciences,Family Studies,KEEP
Social Sciences,Gender Studies,KEEP
Social Sciences,Demography,KEEP
Social Sciences,Social Sciences Methods,KEEP
Social Sciences,Sociology,KEEP
Social Sciences,Law,KEEP
Social Sciences,Science Studies,KEEP
Social Sciences,Information & Library Sciences,KEEP
Social Sciences,Social Work,KEEP

Appendix C

Annotation Guideline

The task proposes the identification of value-chain research phase in scientific outputs, classifying publications and research projects among the research phase of the records, choosing between: **(1) basic research, (2) translational research, (3) clinical research or (4) public health**.

C.1 Category definition

According to this, each record must be categorized in one of the following categories:

1) Basic research (also called fundamental research)

This focuses on discoveries and knowledge, driven by hypotheses that advance the understanding of the unknown; it builds new knowledge; in biomedical sciences it uses cells and model organisms and very rarely human subjects or human biological material. It involves scientific exploration that can reveal fundamental mechanisms of biology, disease or behaviour. Every stage of the translational research spectrum builds upon and informs basic research. It studies the core building blocks of life (such as: DNA, cells, proteins, molecules, etc.) in order to answer fundamental questions about their structures and how they work. For example, oncologists now know that mutations in DNA enable the unchecked growth of cells in cancer. A scientist conducting basic research might ask: How does DNA work in a healthy cell? How do mutations occur? Where along the DNA sequence do mutations happen? And why?

The following topics should be considered part of basic research:

- Tissue, Cellular & Molecular basis of disease
- Tissue, Cellular & Molecular understanding of mechanisms
- Use of Animal models - zebrafish, rats, human cells, fly, c.elegans, mice, rabbit, guinea pig
- Development of techniques - protein, chemistry, molecular, cellular

Some examples of publications in the category:

- *GPR40 full agonism exerts feeding suppression and weight loss through afferent vagal nerve.*
- *The Functional Mammalian CRES (Cystatin-Related Epididymal Spermatogenic) Amyloid is Antiparallel β-Sheet Rich and Forms a Metastable Oligomer During Assembly.*
- *Viral FLIP blocks Caspase-8 driven apoptosis in the gut in vivo.*
- *Apelin enhances the osteogenic differentiation of human bone marrow mesenchymal stem cells partly through Wnt/β-catenin signaling pathway.*
- *Improved yellow-green split fluorescent proteins for protein labeling and signal amplification.*

The following topics should be considered part of basic research:

- Tissue, Cellular & Molecular basis of disease
- Tissue, Cellular & Molecular understanding of mechanisms

- Use of Animal models- zebrafish, rats, human cells, fly, c.elegans, mice, rabbit, guinea pig
- Development of techniques- protein, chemistry, molecular, cellular

Some examples of publications in the category:

- *GPR40 full agonism exerts feeding suppression and weight loss through afferent vagal nerve.*
- *The Functional Mammalian CRES (Cystatin-Related Epididymal Spermatogenic) Amyloid is Antiparallel β-Sheet Rich and Forms a Metastable Oligomer During Assembly.*
- *Viral FLIP blocks Caspase-8 driven apoptosis in the gut in vivo.*
- *Apelin enhances the osteogenic differentiation of human bone marrow mesenchymal stem cells partly through Wnt/β-catenin signaling pathway.*
- *Improved yellow-green split fluorescent proteins for protein labeling and signal amplification.*

2) Translational research (also called pre-clinical research)

This focuses on translating the discoveries into usability in the clinic, uses large scale testing and both animal models and human biological material. There is a focus on applicability. It connects the basic science of disease with human medicine. During this stage, scientists develop model interventions to further understand the basis of a disease or disorder and find ways to treat it. Testing is carried out using cell or animal models of disease; samples of human or animal tissues; or computer-assisted simulations of drug, device

or diagnostic interactions within living systems. For this area of research the end point is the production of a promising new treatment that can be used clinically or commercialized ("brought to market"). This enterprise is vital, and has been characterized as follows: "effective translation of the new knowledge, mechanisms, and techniques generated by advances in basic science research into new approaches for prevention, diagnosis, and treatment of disease is essential for improving health."

The following topics should be considered part of translational research:

- Study of processes or diseases with the intent to treat
- Drug and vehicle development (since they have a therapeutic target)
- Pre-clinical models (even advanced ones like sheep and pigs)
- With patients samples only as proof of concept, as in tumour samples/biobank usage which is not central to the paper
- With patients samples to establish research pre-clinical models (like in cell lines)

Some examples of publications in the category:

- *Characterization of a porcine model of atrial arrhythmogenicity in the context of ischaemic heart failure.*
- *Assessment of an ultrasound-guided technique for catheterization of the caudal thoracic paravertebral space in dog cadavers.*
- *Nerve Repair and Orthodromic and Antidromic Nerve Grafts: An Experimental Comparative Study in Rabbit.*
- *Murine SIGNR1 (CD209b) Contributes to the Clearance of Uropathogenic Escherichia coli During Urinary Tract Infections.*

- *Notopterol-induced apoptosis and differentiation in human acute myeloid leukemia HL-60 cells.*

3) Clinical research

This searches by testing a specific treatment or procedure, drug, diagnostic or any technology on patients, focusing not only on the biological mechanisms (if applicable) but also on issues of safety, delivery and protocols for implementation. This is the stage of research where clinical trials tend to take place. It includes studies to better understand a disease in humans and relate this knowledge to findings in cell or animal models, testing and refinement of new technologies in people, testing of interventions for safety and effectiveness in those with or without the disease, behavioural and observational studies, and outcomes and health services research. The goal of many clinical trials is to obtain data to support regulatory approval for an intervention. It explores whether new treatments, medications and diagnostic techniques are safe and effective in patients. Physicians administer these to patients in rigorously controlled clinical trials, so that they can accurately and precisely monitor patients' progress and evaluate the treatment's efficacy, or measurable benefit.

The following topics should be considered part of clinical research:

- Clinical trials
- Research regarding patients treatment protocol
- Research implicating patients directly
- Research with patient samples as central feature (genetics of disease, biomarkers, prognostic markers,...)

- Diagnostic of disease
- Classic Epidemiology- cohorts
- Psychiatry (Mental disorders)
- Healthcare standards and guidelines

Some examples of publications in the category:

- *Characterization of a porcine model of atrial arrhythmogenicity in the context of ischaemic heart failure.*
- *Supraclavicular versus infraclavicular approach in inserting totally implantable central venous access for cancer therapy: A comparative retrospective study.*
- *The effect of apolipoprotein E polymorphism on serum metabolome - a population-based 10-year follow-up study.*
- *Functional variations of the TLR4 gene in association with chronic obstructive pulmonary disease and pulmonary tuberculosis.*
- *Comparing patterns of volatile organic compounds exhaled in breath after consumption of two infant formulae with a different lipid structure: a randomized trial.*

4) Public health

This is defined as "the art and science of preventing disease, prolonging life and promoting health through the organized efforts of society" (Acheson, 1988; WHO)**?**. Activities to strengthen public health capacities and service aim to provide conditions under which people can stay healthy, improve their

health and wellbeing, or prevent the deterioration of their health. Public health focuses on the entire spectrum of health and wellbeing, not only the eradication of particular diseases. Many activities are targeted at populations such as health campaigns. Public health services also include the provision of personal services to individual persons, such as vaccinations, behavioural counselling, or health advice.

The following topics should be considered part of public health:

- Clinical trials
- Research regarding patients treatment protocol
- Research implicating patients directly
- Research with patient samples as central feature (genetics of disease, biomarkers, prognostic markers,...)
- Cultural/socioeconomic impact on Health
- Health Policy
- Global Health
- Population Health

Some examples of publications in the category:

- *Post-elimination surveillance in formerly onchocerciasis endemic focus in Southern Mexico.*
- *Association of intestinal colonization of ESBL-producing Enterobacteriaceae in poultry slaughterhouse workers with occupational exposure-A German pilot study.*

- *Use of non-HIV medication among people living with HIV and receiving antiretroviral treatment in Côte d'Ivoire, West Africa: A cross-sectional study.*
- *How did the use of psychotropic drugs change during the Great Recession in Portugal? A follow-up to the National Mental Health Survey*
- *Prevalence and social burden of active chronic low back pain in the adult Portuguese population: results from a national survey*

C.2 Annotation process

The annotations process must be done by following different steps. The next lines describe the step-by-step process that the annotator must accomplish for each publication or R&D project:

1. Read title and abstract of the record. If the text is not in English, or is not in the biomedical domain, label it as `Discarded`.

2. Assign the category basic/translational/clinical/public health research according to the research phase it takes part of (if in doubt, go on to the next one, and you could add the category Doubt) a. Assign the record one of the following labels: - basic - translational - clinical - public health - In case of doubt, you can add the label doubt. In this case, the agreement will be completed with the other two annotators, or it will be discussed later. b. Finish the labelling process for the publications and projects, continue with the next record.