Un Método para la Detección de Controversia en Textos y su Aplicación al Caso de Comentarios sobre Fármacos en Foros de Salud

Trabajo Fin de Máster

Ezequiel López López

Trabajo de investigación para el

Máster en Tecnologías del Lenguaje

Universidad Nacional de Educación a Distancia

Dirigido por:

Prof. Dr. D. Jorge Carrillo de Albornoz Cuadrado

Prof. Dra. Dña. Laura Plaza Morales

Septiembre 2020

Agradecimientos

Este trabajo no hubiera sido posible, en primer lugar, sin la existencia de una institución de enseñanza a distancia con todas la garantías de la educación pública en el marco europeo, como es la UNED. Tampoco podemos olvidar la educación pública, tan importante, que nos ha traído hasta aquí y que debe seguir trayendo a mucha gente.

Dentro de la enormidad de la UNED, quiero agradecer a mis directores, Jorge Carrillo de Albornoz y Laura Plaza Morales, su extraordinaria labor de dirección, su pasión por este campo y su actitud, que hacen sentirse bienvenidas a todas las ideas nuevas que van surgiendo en el proceso. Igualmente a los creadores, coordinadores y directores de programas de máster como este, que nos preparan para un futuro donde la interdisciplinaridad estará presente en todas las áreas, especialmente en la de las tecnologías de la información.

He de agredecer a IBM por haberme descubierto este campo, que ha resultado ser vocacional. Y a Deutsche Bahn, por darme la oportunidad de seguir profundizando en él, y aplicarlo en la industria.

Por último, agradecer el apoyo moral a familia y amigos, en la distancia y en la cercanía. Y especialmente a Sonja Sudimac, que ha vivido este trabajo tanto como yo.

Resumen

La controversia, como fenómeno social y lingüístico, consiste en la discusión o debate reiterado de individuos con posiciones enfrentadas. En la actualidad, goza de una especial visibilidad gracias a las condiciones idóneas de una sociedad hiperconectada, que han permitido registrar y potenciar la interacción de usuarios *online*, a menudo anónima, así como la creación y consumo de contenido nunca antes visto.

Analizar las propiedades y características propias de este fenómeno puede permitirnos extraer diferentes *insights* sobre el tema que es objeto de controversia: un mejor entendimiento del porqué de su controversia, su percepción en la comunidad, si el fenómeno de controversia es equivalente para diferentes dominios y facilitar el desarrollo de herramientas que mejoren el acceso y consumo de la información para los usuarios, entre otros aspectos de interés. Sin embargo, debido a su sutileza y dependencia del contexto, su definición y detección es aún un paradigma sin resolver.

En este trabajo se ha realizado un estudio del problema de la detección de controversia en textos, identificando cuáles son los desafíos de las metodologías existentes en el estado del arte para este problema. Entre estos desafíos, encontramos una falta de definición explícita y ampliamente aceptada y aplicada, así como una metodología para su detección acordemente amplia e independiente del dominio y caso de uso. Para afrontar dichos desafíos, hemos desarrollado una propuesta para una definición amplia de controversia, independiente del dominio, y una aproximación técnica para su detección, además de su implementación y evaluación en un caso de estudio concreto: el de comentarios de usuarios en foros del ámbito médico (corpus *Drug Review Dataset*).

Dicha propuesta se ha basado, por un lado, en la novedosa aplicación formal de detección de argumentación como base para la detección de controversia, y por otro lado, incluyendo otros aspectos presentes en el estado del arte, como son la formación de grupos de opinión y la confrontación de dichos grupos respecto al tema de controversia.

Se ha desarrollado un sistema modular de detección basado en dicha definición, consistente en un detector de argumentos, un componente de *clustering* de argumentos, un clasificador de polaridad y un estimador de controversia, de propuesta

propia. Para dicho componente, se han conseguido resultados de clasificación de argumentos que superan los encontrados en el estado del arte para el mismo problema y configuración.

Finalmente, hemos evaluado el caso particular *Drug Review Dataset*, comparando los resultados con una anotación manual para el mismo dataset, llevada a cabo por tres anotadores diferentes. Los resultados obtenidos son prometedores, detectando la controversia correctamente en sus extremos y aportando una serie de detalles para su explicabilidad.

Abstract

Controversy, as a social and linguistic phenomenon, consists of repeated discussion or debate by individuals with opposing positions. Nowadays, it is highly visible thanks to the ideal conditions of a hyperconnected society. This has allowed to record and increase user interaction *online*, often anonymous, as well as the creation and consumption of content never before seen.

Analyzing the properties and characteristics of this phenomenon can allow us to extract different *insights* on the subject that is the subject of controversy: a better understanding of why it is controversial, its perception in the community, if the phenomenon of controversy is equivalent for different domains and facilitate the development of tools that improve access and consumption of information for users, among other aspects of interest. However, due to its subtlety and dependence on the context, its definition and detection is still an unresolved paradigm.

In this work, we study the problem of controversy detection in texts, identifying what are the challenges of the existing methodologies in the state of art for this problem. Among these challenges, we find a lack of explicit and widely accepted and applied definition, as well as a methodology for detection that is consistently broad and independent of the domain and case of use. To address these challenges, we have developed a proposal for a broad, domain-independent definition of controversy and a technical approach for its detection, as well as its implementation and evaluation in a specific case study: user comments in medical forums (corpus *Drug Review Dataset*).

This proposal is based, firstly, on the novel formal application of argument detection as a basis for the detection of controversy, and secondly, on including other aspects present in the state of art, such as the formation of opinion groups and the confrontation of these groups with respect to the subject of controversy.

A modular detection system has been developed based on this definition, consisting of an argument-detector, a component for argument *clustering*, a polarity classifier and controversy estimator proposed in this work. For this component, obtained argument classification results exceed those found in the state of art for the same problem and configuration.

Finally, we have evaluated the particular case of *Drug Review Dataset*, comparing the results with a manual annotation for the same dataset, carried out by three different evaluators. The results obtained are promising, detecting the controversy correctly in its extremes and providing a series of details for its explanation.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

1.1. Motivación

La **controversia** es un fenómeno social y lingüístico, fruto del intercambio de información entre individuos, presente en nuestro día a día, en forma de discusión reiterada de opiniones contrapuestas entre diferentes individuos o grupos de individuos.

Hoy en día, la controversia goza de una especial visibilidad gracias a las condiciones idóneas de una sociedad hiperconectada, que han permitido registrar y potenciar esta interacción *online* y por tanto, también dicha controversia. Además de esta gran conectividad y ritmo de interacción, el carácter anónimo de esta interacción, en gran medida, ha fomentado que los usuarios generen contenido de una manera no contemplada con anterioridad.

Debido a sus características lingüísticas y su carácter sutil y contextual, la detección de controversia es un problema particularmente interesante y complejo en el campo del Procesamiento del Lenguaje Natural y constituye el objeto de estudio de este trabajo, tomando como referencia el caso de comentarios de usuarios sobre fármacos en foros del ámbito médico.

Analizar las propiedades y características propias de este fenómeno puede permitirnos extraer diferentes *insights* sobre el tema que es objeto de controversia, un mejor entendimiento del porqué de su controversia, su percepción en la comunidad, si el fenómeno de controversia es equivalente para diferentes dominios y facilitar el desarrollo de herramientas que mejoren el acceso y consumo de la información para los usuarios, entre otros aspectos de interés.

En este caso particular de estudio, podemos encontrar una enorme utilidad en analizar aquellos fármacos que generan controversia y así tener una mejor comprensión de los hechos que existen detrás de estas posiciones contrapuestas.

Esta información es de enorme valor en el mundo farmacológico, a la hora de validar, reportar o mejorar fármacos, ya que aporta una nueva visión, no tan presente

actualmente: el uso, efectos y adopción reales de fármacos en la población a nivel global. También podría utilizarse para detectar si existen sesgos en la percepción o exposición de la información en algunas de las partes.

A nivel general, una aproximación como la presentada en este trabajo podría ser extrapolada a otros ámbitos, ya que la propuesta en sí es de carácter generalista, concretada para el caso de uso específico mediante corpora e indicadores de dicho dominio.

Un sistema y una medida de este tipo, puede constituir una herramienta novedosa que aportar al estudio de la desinformación en Internet, la evaluación de calidad de contenido y un nuevo indicador para los usuarios a la hora de considerar la veracidad de cierta información en Internet, mejorando el acceso y uso de la información para los usuarios tanto dentro como fuera de la Web.

Extensión del uso de Internet

En las últimas décadas, Internet se ha convertido en un medio insustituible para el acceso a la información. Acceder, publicar, editar y en definitiva consumir información se ha convertido en un actividad que realizamos de manera natural en nuestro día a día y desde cualquier lugar.

Esta amplia adopción ha sido el resultado de años de desarrollo tecnológico en forma de dispositivos más potentes, mejores infraestructuras de comunicación y de la naturaleza intrínsicamente colectiva de Internet, que ha incentivado que muchas actividades colectivas que tenían lugar sólo en el mundo *offline* se hayan desplazado y potenciado en el mundo *online*.

Este desarrollo ha provocado una democratización en el uso de Internet, tanto de consumo como de creación de contenido. El perfil de usuario ha pasado de ser un perfil mayoritariamente técnico, ligado al mundo de la investigación y la tecnología, a ser un usuario universal en todos los aspectos (edad, clase social, sexo, ideología, nacionalidad, etc).

Con la evolución del perfil de usuario, también ha cambiado la forma de interaccionar con la información, así como el tipo, calidad y cantidad de información que se genera. Hemos pasado de un contenido de carácter práctico, informativo, científico y oficial, a un escenario donde ese contenido convive con experiencias personales que se comparten, a menudo, sin ningún objetivo particular.

En este nuevo escenario de nuevos usuarios, nuevas plataformas de intercambio de información, y debido al ritmo con el que nuevo contenido es creado, propagado y consumido, resulta un verdadero desafío realizar una evaluación del contenido que se publica. Esto se traduce en la práctica en la imposibilidad de distinguir el contenido proveniente de la experiencia de un usuario de las informaciones que refieren a un evento o hecho que es descrito de manera objetiva con la intención de informar.

Consumo de información y subjetividad

La información compartida por una comunidad anónima de usuarios es considerada a menudo como veraz por la comunidad, sin que pase necesariamente a través de ningún proceso de medida de calidad de contenido o verificación. De igual forma, el concepto de reputación en dichas comunidades surge de sistemas relativamente rudimentarios basados en criterios cuantitativos como número de *likes* o número de publicaciones que asignan autoridad a diferentes autores de contenido dentro de la comunidad sin que represente un grado de la objetividad, *expertise* o veracidad acerca de su contenido.

A pesar de la subjetividad inherente a estas publicaciones, el contenido ligado a la percepción subjetiva de un usuario puede ser de gran utilidad o de gran influencia para otros usuarios. De esta manera, es posible para un usuario, por ejemplo, considerar experiencias de otros usuarios a la hora de tomar una decisión respecto a la compra de un producto, visita de un restaurante o uso de un medicamento, entre otros. Podríamos esperar que, bajo la asunción de que cada pieza de información constituirá una visión sesgada y subjetiva, a medida que consideramos nuevas piezas de información estaríamos aumentando el grado de objetividad sobre el tema en cuestión. Sin embargo, no existe siempre un principio objetivo y factual para cualquier tema, como es el caso de temas ideológicos o éticos, por ejemplo. Incluso dándose este principio factual, existen múltiples razones por las que puede percibirse de una manera muy diferente entre usuarios y generar posiciones contrapuestas.

A este conjunto de temas, susceptibles de no alcanzar un consenso en una comunidad de individuos, los identificamos como controvertidos. Consecuentemente, llamaremos **controversia a la situación generada, de debate reiterativo de posiciones contrapuestas entre individuos o grupos de individuos**.

Controversia

La controversia es un fenómeno muy común en nuestra sociedad, que con la llegada y expansión de Internet ha ganado una nueva dimensión y visibilidad, al alcanzar un carácter global y sin límites geográficos o temporales.

En estas circunstancias, pueden darse múltiples situaciones en las que puede generarse controversia. Hemos comentado ya el impacto que ha tenido el poder compartir, como usuarios, nuestra visión con el resto de la red y que esto puede alimentar controversias en diferentes ámbitos. Pero también existen otras vías como, por ejemplo, el acceso a información más detallada sobre hechos del pasado, que pasan hoy a través de un proceso revisionista como es el caso del tráfico de esclavos o la minimización de roles femeninos en eventos importantes de las diferentes áreas del conocimiento.

Los motivos por los que un tema es considerado controvertido son de diversa

índole, ligados normalmente a la información que dispone el opinador, y las herramientas que usa sobre dicha información para construir su postura, así como su contexto. Una misma información, en dos contextos diferentes puede llevar a los usuarios a adoptar dos posiciones contrarias.

Encontraremos pues, que un tema como la *esclavitud* era controvertido a mediados del siglo XIX en Estados Unidos, pero no lo es en la actualidad. Al mismo tiempo que la pena de muerte lo sigue siendo en algunos estados, frente al consenso durante gran parte del siglo pasado.

En otros casos, encontramos un tema presentado de manera sesgada, como es el caso de los productos farmacéuticos, los cuales son presentados comercialmente como *efectivos*, sin que se especifique siempre exactamente en qué casos o bajo qué circunstancias tienen dicha efectividad. Dichas afirmaciones se remiten normalmente a bases estadísticas, que pueden estar o presentarse, a su vez, de manera sesgada. La expectativa generada sobre un producto farmacéutico, unida a la amplia casuística de síntomas y diferentes características de cada usuario, desemboca en ocasiones en un escenario de controversia en el que los diferentes usuarios presentan argumentaciones a favor o en contra del uso del fármaco o de alguno de sus aspectos.

Este escenario es particularmente del interés de este trabajo, en el que se propondrá e implementará una aproximación para detectar el grado de controversia generada en el ámbito de los productos farmacéuticos, sirviéndonos para ello de comentarios de usuarios sobre tales productos, recopilados a partir de foros médicos y redes sociales.

1.2. Definición del problema

Existen diversas manifestaciones de la controversia en las diferentes plataformas donde interaccionan los usuarios, como es el ejemplo de *Wikipedia*, donde encontraremos artículos que son editados y re-editados una y otra vez bajo diferentes matizaciones de los usuarios. En este caso, el número de *contra-ediciones* del artículo puede considerarse como una medida cuantitativa sobre el grado de controversia que genera un tema en la comunidad.

Sin embargo, en la mayoría de los casos, esta estimación no resulta tan trivial. Ante la actual falta de corpora ya etiquetados como controvertidos y no controvertidos, nuestro primer paso deberá ser derivar una definición apropiada para controversia y a partir de ahí, extraer las características adecuadas para su detección y posterior evaluación.

Encontrar y aplicar una definición amplia de controversia que sea válida para diferentes casos, sin basarnos enteramente en características y datos disponibles para casos particulares, es aún un problema sin resolver, como se discute en (Dori-Hacohen, Yom-Tov, y Allan, 2015). Por lo tanto, debe constituir el primer paso en

una aproximación como la presentada en este trabajo y uno de nuestros primeros objetivos.

En la literatura encontraremos definiciones implícitas de la controversia, dependientes de las metodologías y los datos que se usan en cada aproximación. No obstante, no se encuentra una definición explícita o formal, que pueda ser a continuación aplicada en casos particulares, adaptando una metodología general, independiente del contexto, ámbito y fuente de los datos.

Sin embargo, podemos encontrar algunos elementos comunes que podemos considerar para construir una definición que sea lo más amplia posible, sin que deje de ser técnicamente aplicable.

Entre estos elementos, podemos extraer y concretar que la controversia debe contener **confrontación**, por tanto deben existir grupos distinguibles y relevantes en tamaño que tengan parte en esta confrontación. La confrontación en sí misma debe ser detectable, al menos en su **polaridad**, y al mismo tiempo, debe tener cierta base **argumental** que nos permita distinguir una confrontación fundamentada frente a un simple discusión polarizada o un conjunto de opiniones irracionales.

Esta última componente de argumentación no tiene gran presencia en las propuestas de detección de controversia del estado del arte actual, como veremos a continuación, pero resulta un aspecto clave en nuestro enfoque y es la base sobre la que construiremos el resto del proceso.

Detectar la controversia

Como consecuencia de la falta de una definición amplia de controversia, aceptada y utilizada por la comunidad, encontraremos que tampoco existe una metodología ampliamente aceptada para su detección y es el segundo punto que nos proponemos desarrollar en este trabajo.

Sin embargo, podemos encontrar diferentes aproximaciones que han sido desarrolladas para afrontar el problema de la detección de controversia, normalmente ligadas a un tipo de contenido o a un ámbito en particular.

Por una parte encontramos aquellas que son muy dependientes de las características del corpus considerado, como puede ser el caso de Wikipedia, que ofrece algunos datos interesantes como número de ediciones y correcciones de un artículo y una clasificación previa de contenido controvertido (Bykau et al., 2015). Encontramos un caso parecido en aquellas aproximaciones que utilizan datos de *Twitter*, con características propias como los *tweets, menciones, réplicas, etc.* Estas aproximaciones proponen una definición de controversia que es prácticamente intrínseca de la plataforma que se considere y que reduce enormemente las posibilidades de ser extrapolada a otros casos (Popescu y Pennacchiotti, 2010)(Garimella et al., 2018).

En otros casos, encontramos enfoques basados completamente en el análisis de sentimiento (Wang y Cardie, 2016), en los que se realiza una estimación de

la polaridad para el tema dado, basado en el cojunto de elementos de texto (por ejemplo, *tweets*) y se identifica implícitamente controversia con alta polaridad. Este enfoque puede resultar demasiado simplista, ya que podemos estar ante un escenario muy polarizado sin que exista realmente controversia sobre el tema (por ejemplo, aficionados *ultras* de dos equipos de fútbol *twitteando* antes de un partido).

Otro tipo de aproximación, es aquella basada en las características semánticas de la controversia, en la que partiendo de un corpus previamente anotado (de nuevo, extraído de Wikipedia), se realiza una representación vectorial mediante *word embeddings* y la construcción de un modelo de clasificación basado en redes neuronales (Linmans, van de Velde, y Kanoulas, 2018). Los resultados para este tipo de aproximación mejoran los métodos mencionados anteriormente, pero siguen manteniendo una dependencia del ámbito y dataset utilizado. Tampoco se define explícitamente lo que es controversia, lo cual depende de la definición implícita en el corpus de *Wikipedia*, y encontrando dificultades para seguir minando las razones detrás de dicha controversia y la explicabilidad en general de los resultados.

Encontramos una aproximación interesante en (Garimella et al., 2018), basada en grafos, que supera muchos de los inconvenientes que encontramos en las demás aproximaciones y tiene la capacidad de aportar mucha información sobre las razones detrás de la controversia y que además, tiene como objetivo la cuantificación de esta, algo novedoso con respecto al resto. Pero al igual que en otros trabajos mencionados, encontramos una definición de controversia ligada a la metodología para un caso de uso particular. En este caso, la controversia se define como un cálculo de probabilidades de *Random Walk* por un grafo que representa las conexiones entre usuarios que interaccionan en una red, y que representa el grado de particionamiento presente en la red. A mayor particionamiento de grafo, mayor controversia.

Al estar basado en grafos, este enfoque tiene gran relevacia para casos en los que se da una estructura de relaciones clara, a veces ya dada como en *Twitter*, donde tenemos la información sobre las interacciones explícitamente. Sin embargo, se complica cuando queremos aplicarlo de manera amplia a otros casos, donde estas estructuras son muy sutiles o no vienen determinadas explícitamente y dichos grafos deben inferirse.

Una propuesta que no tiene gran presencia en el estado del arte, como hemos comentado anteriormente, es el uso de *argument mining* como una posible vía para atacar el problema de la detección de controversia, como se describe en (Addawood y Bashir, 2016) y (Dori-Hacohen, Yom-Tov, y Allan, 2015), donde se apunta un cierto paralelismo entre el problema de detección de controversia y el de detección de argumentos, pero no llega formalmente a desarrollar cómo sería tal aproximación.

Encontramos particularmente útil este concepto, ya que en muchas de las aproximaciones descritas se precisa de un primer paso que aporte consistencia y sirva de hilo conductor del proceso de detección. En el caso de (Garimella et al., 2018),

este fundamento esta cubierto por las aportaciones del enfoque basado en grafos, que analiza las estructuras internas de la controversia en una manera similar a la que se usa normalmente en el campo de *argument mining*.

Nuestra definición de controversia se basa en la confrontación de diferentes grupos temáticos o de opinión, de tamaño relevante y comparable, con polaridades opuestas. La estructura de dicha confrontación o debate encaja perfectamente dentro de las estructuras argumentativas consideradas en la minería de argumentos, ya que tendremos diferentes individuos exponiendo razones para apoyar o atacar un tema de controversia, a través de sus diferentes aspectos o sub-temáticas.

El problema de la detección de argumentos es un problema mucho más definido, estudiado y acotado que el de la detección de controversia. Si enmarcamos el problema de la detección de la controversia como un *subtipo* del problema de detección de argumentos, tendremos una vía de desarrollo mucho más amplia y con más garantías, por el conocimiento que existe sobre este marco.

Partiendo de este escenario en el que los elementos de texto considerados deben contener indicios de argumentación, podemos completar las demás condiciones necesarias para la controversia, la formación de grupos de opinión y el análisis del sentimiento y polaridad, de marcada relevancia en el estado del arte.

Una nueva aproximación

Partiendo de nuestra propuesta de definición para la controversia que incluye las condiciones de presencia de argumentación y existencia de grupos de opinión suficientemente grandes junto a la condición de polaridad que manifiesta la confrontación de estos grupos, podemos afrontar el segundo de nuestros objetivos, la propuesta de un *pipeline* de detección de controversia.

En dicho *pipeline* encontraremos una serie de pasos con la funcionalidad de analizar las diferentes condiciones que hemos impuesto en la definición de controversia. Es decir, encontraremos un *pipeline*, como se describe en el capítulo de Sistema (3), con los siguientes pasos: *(i)* preprocesamiento usual y segmentación en unidades de texto, *(ii)* detección de argumentación mediante la construcción de un sistema de clasificación, *(iii)* clusterización de los elementos argumentativos, *(iv)* análisis de polaridad de los *clusters* obtenidos, *(v)* estimación de controversia.

Esta propuesta es de carácter general, no dependiente del ámbito, contenido o plataforma sobre la que se aplique (salvo, por supuesto, los pasos de preprocesado), algo que no hemos encontrado en el estado del arte. Hemos utilizado como caso de aplicación, el caso de uso de comentarios de usuarios sobre fármacos en foros del ámbito médico y redes sociales (Gräßer et al., 2018), tratando cada comentario de manera anónima e individual, donde el objeto de estudio de controversia que consideramos es un caso de uso (uso del fármaco para una sintomatología concreta: ej. paracetamol-cefaleas). El proceso de segmentación de texto se ha planteado

para un sistema basado en procesamiento de oraciones, aunque podría trabajarse fácilmente con diferente granularidad.

En el paso de detección de argumentos, hemos utilizado una serie de features (Stab y Gurevych, 2017) presentadas en el estado del arte para construir un clasificador sobre un dataset multi-dominio público con 25,000 instancias de oraciones anotadas como argumento / no-argumento (Stab et al., 2018), igualando los resultados obtenidos por dichos trabajos.

Este clasificador es de carácter genérico y ha sido complementado con otras dos señales propias del dominio sobre el que hemos desarrollado la adaptación del *pipeline*, siendo este el único punto del proceso en el que introducimos elementos del dominio. En nuestro caso, un segundo clasificador basado en *features* del ámbito médico (presencia de síntomas, posologías, unidades de medida, efectos secundarios, etc.) con resultados de clasificación de $F1 = 0{,}78$ sobre el set de evaluación y separadamente una señal sobre la presencia de síntomas, de gran relevancia, extraída gracias a la ontología médica de *UMLS*[1] (Bodenreider, 2004).

Finalmente hemos combinado la señal del clasificador genérico, la señal del clasificador de dominio y la señal de síntomas mediante voto mayoritario, obteniendo resultados de clasificación de $F1 = 0{,}68$ sobre una muestra aleatoria de 215 elementos etiquetados manualmente.

Se ha realizado a continuación un proceso de clustering aglomerativo jerárquico, utilizando las representaciones vectoriales de las oraciones argumentativas mediante *word2vec* y calculando su similitud mediante la distancia de coseno (Santus et al., 2018) . De esta manera, podemos agrupar semánticamente los temas sobre los que tratan los argumentos. Para los clúster, analizaremos cuál es su sentimiento neto (positivo o negativo), a través de un clasificador del estado del arte (Tai, Socher, y Manning, 2015), usando una red LSTM y la representación de *word embeddings* de las oraciones argumentativas.

A partir de los detalles obtenidos, aplicaremos la definición de controversia sirviéndonos de la media de las polaridades netas de los *clusters* pesadas por su tamaño y proporcionando así una estimación de la controversia para cada caso de uso de fármacos.

Para su evaluación, hemos seleccionado aleatoriamente 10 casos de uso diferentes, se han establecido unos criterios de evaluación por los que se identificará como *no-controvertido* aquellos casos que claramente no lo son, y *controvertido* en caso contrario, siendo anotados independientemente por tres evaluadores, con un resultado de acuerdo del 86 % y una $\kappa = 0{,}73$.

Debido a la naturaleza sutil y subjetiva de la controversia, se ha realizado finalmente un análisis cualitativo que arroje algo de luz sobre qué aspectos han influido en que los diferentes casos se hayan identificado como controvertidos o no controvertidos. Este análisis aportará un mayor significado a los resultados que un

[1] Unified Medical Language System

análisis cuantitativo, debido a la limitación de tiempo y recursos para un análisis de estas características.

Como conclusiones, exponemos la viabilidad de una propuesta de estas características y también sus limitaciones, aportando como argumento a favor el hecho de obtener resultados prometedores tras haber realizados estos experimentos sobre el prototipo del sistema. Cabe recordar que cada paso del *pipeline* tiene un gran margen de mejora, ya que no nos hemos detenido de manera exhaustiva a optimizar sus componentes individuales, sino en desarrollar la propuesta del *pipeline* completo, inferido a partir de la definición de controversia propuesta.

1.3. Propuesta y objetivos

Partiendo del estado del arte con respecto a la detección de controversia en el lenguaje, trataremos primeramente de proponer una definición de controversia amplia sobre la que desarrollar el resto del trabajo. Una vez definida y modelada nuestra definición de controversia, analizaremos las distintas técnicas de Procesamiento de Lenguaje Natural adecuadas para el caso, tales como *clustering*, técnicas de clasificación de texto y *sentiment analysis*, entre otras. Como caso de estudio, hemos determinado que el entorno médico es un excelente punto de partida para probar nuestra definición y técnicas propuestas. En concreto, el análisis de redes sociales sobre el uso de fármacos, un tema controvertido cargado de subjetividad y experiencias de usuarios.

Nuestra hipótesis consistirá en que los comentarios de usuarios sobre el uso de fármacos contienen información suficiente para estimar el grado de controversia que un fármaco concreto genera en la población de usuarios y que dicha controversia puede ser estimada a través de técnicas de procesamiento de lenguaje natural. Realizaremos las siguientes asunciones para el desarrollo de dicha hipótesis:

- Puede desarrollarse una definición amplia de controversia suficiente para establecer la base del desarrollo de nuestro método de estimación.
- Para que dicha controversia se manifieste, deben de existir grupos de opinión enfrentados de tamaños relevantes y comparables, donde se expongan de manera argumentada diferentes experiencias positivas o negativas sobre el uso del fármaco.
- Es posible detectar la argumentación en dichos comentarios, a través del uso de diversos indicadores que debemos definir e identificar, algunos propios del contexto y otros generales, y mediante la selección y uso de corpora ya anotados en este aspecto.
- Existen indicadores generales de argumentación, y por tanto, podemos utilizar hasta cierto punto, corpora anotados en argumentación pertenecientes a dominios diferentes al nuestro.

- Los grupos de opinión necesarios para representar las posiciones enfrentadas propias de nuestro concepto de controversia son identificables mediante técnicas de agrupamiento tales como *clustering*.
- El enfrentamiento de dichas posiciones puede ser estimado mediante una evaluación de sentimiento, a través de técnicas de *Sentiment Analysis* existentes.
- El objeto de la controversia considerado será el caso de uso de un fármaco para un trastorno y estos casos de uso serán comporables entre sí, y evaluables a través de la revisión manual de los comentarios que los componen.

Para su desarrollo, partimos de la base de la existencia de grupos definidos de opinión contrapuesta entre los usuarios de un mismo fármaco. Por tanto, debemos primero identificar dichas opiniones mediante el análisis automatizado de oraciones en dichos textos, para encontrar afirmaciones en una dirección u otra, así como los respectivos argumentos que las apoyan.

Una vez encontradas dichas argumentaciones, podremos analizar si pertenecen a una misma corriente de opinión, y así agruparlas dentro de un mismo tema o argumento. Analizando las temáticas de dichos grupos, mediante los argumentos extraídos de las opiniones, podemos decidir si dichos grupos se encuentran en posiciones de opinión contrapuestas, y si ambos son de un tamaño relevante.

Si encontramos este fenómeno, confirmando nuestra hipótesis inicial, seremos capaces de realizar una estimación automática de controversia para cada fármaco, y habremos definido una metodología que puede ser potencialmente extrapolada a otros casos y dominios.

Para el desarrollo de dicho método, nos marcamos los siguientes objetivos:

- Desarrollo de una definición amplia de controversia, basada en los enfoques analizados a través del estudio del estado del arte.
- Definición de los pasos necesarios para el procesamiento de texto para la identificación de controversia, basándonos en la definición anterior y el estado del arte de las técnicas utilizadas en el campo. Entre estos pasos, podemos distinguir principalmente los siguientes sub-objetivos:
  - Definición e identificación de los indicadores de argumentación para nuestro caso concreto.
  - Estudio de viabilidad de la aplicación *cross-domain* de corporas desarrollados para la detección de argumentación.
  - Exploración y selección de técnicas adecuadas para la agrupación de opiniones para nuestro caso.
  - Exploración y selección de técnicas adecuadas para el análisis de sentimiento o polaridad para nuestro caso.

- Desarrollo de un sistema de procesamiento de lenguaje natural, para la detección de dicha controversia, basado en las definiciones anteriores.
- Aplicación del enfoque generalista propuesto para la identificación de controversia a un caso particular, como es el de comentarios de usuarios de fármacos en redes sociales del ámbito médico, integrando elementos del lenguaje médico y farmacológico.
- Obtención y análisis de resultados, y su potencial validación mediante otros conjuntos de datos relevantes.
- Proposición de aplicaciones futuras de los resultados de un sistema de estas características.

1.4. Estructura del documento

Este trabajo está estructurado por capítulos de la siguiente manera:

Capítulo 1. Introducción. Este capítulo introduce los principales motivos que han llevado a la realización de este trabajo, así como la problemática y el estado actual de la disciplina. Por último, se presentan las diferentes contribuciones del trabajo realizado.

Capítulo 2. Estado del arte. Este capítulo describe en mayor detalle la disciplina que nos ocupa, presentando su origen y su historia hasta el presente. Se muestran las técnicas actuales más utilizadas para resolver las tareas más relevantes del tema abordado, así como sus debilidades.

Capítulo 3. Método para Detección de Controversia. En este capítulo se describe en profundidad nuestro método para la detección de controversia: nuestra propuesta de definición y el sistema para su detección, así como sus componentes y el proceso desarrollado para su definición e implementación.

Capítulo 4. Evaluación y Discusión. Este capítulo describe la metodología utilizada para evaluar la propuesta realizada, a la vez que presenta los resultados obtenidos al evaluar el método propuesto en diferentes tareas y realiza una discusión sobre el análisis cualitativo de dichos resultados.

Capítulo 5. Conclusiones y trabajo futuro. Este capítulo recopila las diferentes conclusiones extraídas del trabajo realizado, y propone algunas líneas de trabajo futuro.

Capítulo 2

Estado del arte

En este capítulo describiremos de manera amplia el marco teórico del campo en el que situamos el problema de la detección de controversia. Comenzamos presentando el contexto social, lingüístico y tecnológico del fenómeno de la controversia, así como una breve mención al caso concreto que utilizaremos para aplicar nuestra propuesta amplia de detección de controversia. A continuación, encontraremos una exposición de las técnicas actuales propuestas y utilizadas en la detección de controversia así como los demás campos que están implicados en nuestra propuesta, también referenciados por el estado del arte: detección de argumentación, *clustering* y análisis de sentimiento.

2.1. Controversia e Información en Internet

Internet se ha convertido en un medio de propagación de información y contenidos de todo tipo, así como de edición, publicación y consumo. Lo que caracteriza a Internet, desde el punto de vista de la creación de la información, frente a otros medios más tradicionales como la televisión, radio, prensa, etc. es su naturaleza intrínsecamente colectiva.

A pesar de que las fuentes oficiales de información, incluyendo los medios tradicionales mencionados, tienen presencia en Internet, la mayor parte del contenido es creado por una comunidad anónima constantemente. Dicho contenido, en la mayoría de los casos, no sigue una directriz particular en forma o en la información que contiene. En la misma dirección, tampoco es sometido a ningún proceso de revisión o edición, exceptuando casos particulares como Wikipedia, donde los usuarios corrigen y complementan constantemente la información publicada por otros usuarios, generándose finalmente una información que, en la mayoría de los casos, podemos entender como cercana a la objetividad.

Con el desarrollo de las redes sociales en la última década, y los medios tecnológicos que lo han permitido, esta producción de contenido ha crecido de manera

WORLD INTERNET USAGE AND POPULATION STATISTICS 2020 Year-Q2 Estimates						
World Regions	Population (2020 Est.)	Population % of World	Internet Users 30 June 2020	Penetration Rate (% Pop.)	Growth 2000-2020	Internet World %
Africa	1,340,598,447	17.2 %	566,138,772	42.2 %	12,441 %	11.7 %
Asia	4,294,516,659	55.1 %	2,525,033,874	58.8 %	2,109 %	52.2 %
Europe	834,995,197	10.7 %	727,848,547	87.2 %	592 %	15.1 %
Latin America / Caribbean	654,287,232	8.4 %	467,817,332	71.5 %	2,489 %	9.7 %
Middle East	260,991,690	3.3 %	184,856,813	70.8 %	5,527 %	3.8 %
North America	368,869,647	4.7 %	332,908,868	90.3 %	208 %	6.9 %
Oceania / Australia	42,690,838	0.5 %	28,917,600	67.7 %	279 %	0.6 %
WORLD TOTAL	7,796,949,710	100.0 %	4,833,521,806	62.0 %	1,239 %	100.0 %

Figura 2.1: Estadísticas sobre el uso de Internet respecto a la población. Destacamos la columna *Growth 2000-2020* que muestra el incremento en el uso de Internet y que muestra como en algunas regiones el crecimiento es de miles por ciento. (Fuente: Internet World Stats)

exponencial (Fig. 2.1) y también ha ganado heterogeneidad en su forma. Hemos pasado de mera información textual en los foros de finales de los años 90s a aplicaciones especializadas para opinar sobre diferentes temáticas, foros de carácter general que incluyen todo tipo de temáticas bien organizadas, diferentes servicios de blogging, microblogging en tiempo real a través de Twitter, noticias a través de *Facebook*, experiencias a través de Instagram, opiniones a través de *TripAdvisor*, vídeos en canales de *Youtube*, etc.

Paralelamente, ha tenido lugar el desarrollo de tecnologías que permiten consumir y producir este contenido desde cualquier lugar en tiempo real, gracias a nuevos dispositivos móviles y al desarrollo de los servicios de datos móviles. Esta combinación de accesibilidad, conectividad y plataformas ha transformado la sociedad en los últimos años, llegando a ser un catalizador de movimientos sociales que sincronizan miles de personas en cuestión de horas, de manera espontánea para formar parte de un evento o una protesta.

Una gran cantidad de la información que se produce consiste en opiniones y experiencias referidas a un hecho o a un objeto sobre el que el autor escribe. Podemos encontrar foros donde los usuarios describen sus experiencias sobre un destino de viaje, vídeos de *YouTube* recomendando y explicando cómo utilizar un producto concreto, opiniones en cualquier portal de compras como *Amazon*, exposición de algún hecho de actualidad de una manera más o menos subjetiva en blogs o *Facebook*, etc.

Idealmente, bajo la filosofía en la que se desarrolló la Web, esta información debería ser, en términos generales, veraz y honesta, sin una intencionalidad concreta y entendiendo que cada información publicada por un usuario puede tener una componente importante de subjetividad. Sin embargo, la realidad es otra y no podemos garantizar que la información que es publicada por un usuario sea veraz y fiel a la realidad, ya sea de manera intencionada o no.

Por otra parte, no todos los temas sobre los que se crea contenido tienen una

base factual, determinista, aceptada u objetiva. Y cada vez son más difusas las fronteras que separan aquella información que está basada en hechos consolidados y aquella que no, o que ha surgido de una interpretación de esos hechos que no puede considerarse como objetiva.

Fruto de la dificultad de realizar dicha distinción, el usuario no siempre tiene en su poder herramientas necesarias que le permitan enriquecer su juicio sobre la información que está consumiendo. Esto puede llevarlo a tomar decisiones que no están basadas en hechos consolidados, ser influenciado por otros y convertirse él mismo en propagador de desinformación, incluso sin ser consciente.

Proporcionar las herramientas adecuadas y conseguir un control y regulación de la calidad del contenido en Internet es uno de los desafíos actuales en el mundo de las tecnologías de la información.

Las técnicas para afrontar estos desafíos, aún bien diseñadas e implementadas, quedan fácilmente desactualizadas en este contexto, por la cantidad de nuevas estrategias, tecnologías y casuísticas que van surgiendo y esquivando cualquier intento de indentificarlas, analizarlas y afrontarlas a tiempo.

Encontramos casos en los que entidades políticas de todo signo utilizan legiones de usuarios ficticios (*bots*) en redes sociales para influenciar la opinión de los usuarios en una u otra dirección. A menudo mediante el uso de información sesgada, falaz o exagerada para atacar a adversarios políticos, simplemente para aumentar las posibilidades de conseguir un mejor resultado electoral, o bien para enmascarar información poco conveniente ante la población.

De la misma manera, encontramos que diferentes marcas de productos generan comentarios positivos a través de usuarios ficticios en diferentes páginas *recolectoras de opinión*, como pueden ser la sección de opiniones para productos en Amazon, eBay, Tripadvisor para restauración y hoteles, Glassdoor para el mundo laboral, o cualquier otro foro especializado en temas específicos como salud.

Como ejemplo reciente, podemos encontrar el caso que más impacto y preocupación ha provocado en los últimos años, las *Fake News* (Lee, 2019). Información fácilmente consumible para el usuario, que parte de un contexto real e introduce una idea que busca provocar una reacción en el usuario. Esta información puede estar basada en una información real o puede ser directamente generada a través de fragmentos de información de otros contextos. Podemos imaginar un contexto de rumor de fraude electoral, en el que se rescata una fotografía de un antiguo escándalo en otro país y se escribe un titular y descripción incendiarios para a continuación propagar dicha información por las redes sociales.

El usuario actual de Internet dista mucho de aquellos primeros usuarios, de perfil mayoritariamente técnico, que usaron y desarrollaron la Web durante las primeras décadas. Hoy encontramos un uso extremadamente extendido de la Web en toda la población, incluyendo perfiles de todo tipo. Podíamos pensar que la consecuencia última de esta democratización del uso de la Web y el consecuente

acceso a la información sería una sociedad más informada y crítica. Sin embargo, este acceso (casi) universal a la información no se ha traducido automáticamente en una sociedad mejor informada, aunque sí más consumidora de información y por tanto, también más susceptible a la manipulación.

La universalización del uso del Internet, del tipo de usuario, la hiperconectividad de los usuarios a través de redes sociales, las estrategias de segmentación de usuarios y algoritmos de recomendación, entre otros factores, ha provocado que muchos usuarios vivan diariamente en una burbuja de información, en la que confían más en la información que llega a través de su red de contactos que en la información que es publicada en medios de comunicación oficiales. Con lo cual también se ha desplazado el concepto de autoridad de las fuentes de información.

En las *burbujas de información* (Pariser, 2011), los usuarios, en su interacción con la Web a través de los algoritmos existentes para búsqueda y recomendación, que nos recomiendan contenido similar a nuestros gustos y patrones actuales, acaban consumiendo solamente información afín a sus gustos, generando un aislamiento de conocimiento para el usuario y no entrando nunca en contacto con otro tipo de información.

Psicológicamente existe una tendencia a preferir contenidos que nos reafirman o que coinciden con nuestros gustos y posiciones, frente a aquellos que nos puedan parecer contrarios a nuestras opiniones, nos den una visión difícil de comprender a priori o que nos fuercen a realizar una reflexión o análisis crítico sobre temas sobre los que ya estamos convencidos en alguna dirección.

Si consideramos un fragmento de información puramente informativo, como puede ser un artículo de Wikipedia, un informe técnico o un artículo de un periódico de prestigio, esperamos que los hechos sean descritos de manera objetiva y racional; entendiendo como objetivo lo más cercano a la realidad del hecho en sí como sea posible, minimizando la opinión, posición y contexto del autor o autores de dicha información. Como podemos imaginar, la objetividad absoluta es difícil de conseguir y en términos filosóficos podríamos discutir siquiera si existe.

Cuando consumimos información que ha surgido de la experiencia del autor, o está expresando su opinión sobre un tema concreto, estamos consumiendo información subjetiva ligada a la percepción particular de su autor.

Por tanto podemos decir que una información es más o menos objetiva, pero no podemos afirmar la cualidad de objetividad de manera absoluta.

En el caso de experiencias y opiniones de usuarios, el aspecto de la racionalidad puede resultar clave para distinguir que información puede sernos de mayor utilidad. Consideramos una opinión como racional si el autor de dicha opinión manifiesta mecanismos lógicos de causalidad, inducción, deducción y similares, aunque estos sean usados de manera incorrecta, puesto que se manifiesta la intencionalidad del autor por desarrollar un razonamiento, ya sea correcto o no. Por el contrario, si una opinión expresa meramente un sentimiento o una impresión que no forma parte

del desarrollo de un proceso de razonamiento, consideraremos dicha opinión como irracional.

Consecuentemente, si queremos conocer la opinión de una comunidad respecto a un tema o producto, estaremos más interesados en leer aquellos comentarios que estén basados en un razonamiento y justificación, frente a otros que simplemente expresan una manera de percibir ese tema. El problema de diferenciar si dichos razonamientos están construidos de manera válida y coherente desde el punto de vista lógico, es un problema mayor.

Las opiniones y experiencias relatadas por un usuario tienen una naturaleza intrínsecamente subjetiva, ligada al autor. Sin embargo, si el autor de la opinión es considerado como una autoridad por la comunidad en dicho tema, por su reputación, entonces entendemos que dicha opinión tiene un grado mayor de objetividad y marca el umbral de lo que la comunidad entiende como verdad oficial (o lo más cercano a ella).

De la misma manera, consideraremos objetivas las noticias publicadas por un periódico aceptado por la comunidad como objetivo, por el hecho de ser referenciado y respetado por diferentes sectores de la sociedad, con diferente signo y posición ideológica.

Tanto el grado de objetividad como la racionalidad de los opinadores constituyen factores claves a la hora de contrastar opiniones respecto a un tema concreto, del que queremos obtener una información lo más veraz posible.

El grado de objetividad, ligado a la reputación y autoridad del autor, no es fácil de extraer generalmente a partir de un conjunto de opiniones. Es habitual encontrar en las diferentes plataformas *recolectoras de opinión* mencionadas (redes sociales, foros, sección de comentarios, etc.) un sistema más o menos rudimentario de reputación para los autores de opinión, normalmente basado en el número de comentarios ya realizados y la valoración de otros usuarios.

Sin embargo, no podemos considerar estos sistemas completamente válidos a la hora de considerar las opiniones de dichos usuarios reputados en ese sistema particular como opiniones más objetivas o más veraces. Tanto la actividad histórica de un autor como el número de usuarios que lo apoyan son criterios insuficientes, ya que desconocemos si el motivo por el que otros usuarios lo consideran una autoridad está de verdad fundamentado y es racional frente a una consonancia irracional. Estaríamos ante una situación de sesgo espontáneo como las mencionadas anteriormente (por ejemplo, *burbuja de información*), en las que nueva información exterior sería difícil de ser considerada jamás por los usuarios.

Debemos por tanto, considerar a todos los opinadores *a priori* como equivalentes y basarnos en los indicadores de racionalidad presentes en sus opiniones como indicadores de que dichas opiniones contienen una información útil (por ejemplo, causal) sobre un tema concreto frente a aquellas opiniones que expresan irracionalmente una posición o una manera de sentir del opinador respecto al tema.

Hemos usado el término *información útil* refiriéndonos al tipo de información que podemos usar para tomar racionalmente una decisión, o para conocer unos hechos a través de diferentes opiniones razonadas de los usuarios. Esto no implica que las opiniones consideradas como irracionales no contengan información, ni que esta información no sea útil para analizar un tema o para tomar decisiones. Podríamos usar simplemente la polaridad de sentimiento que los usuarios manifiestan en sus comentarios respecto a un tema, por ejemplo, una ley o un restaurante. Y podríamos asumir que con una muestra significativa y suficiente de comentarios de usuarios sobre un tema o un producto, estaríamos ante una representación de la realidad de ese producto en el *sentir* de la comunidad, pero no entenderíamos el porqué de dichas posiciones, si estas son justificadas ni hasta que punto el contexto personal de cada usuario afecta a su criterio.

Para superar estas dificultades, podemos considerar que los razonamientos de los diferentes usuarios son comparables y están, hasta cierto punto, libres de la influencia de información que no está presente en la opinión misma.

Es decir, consideraremos que un comentario como *No me gustó la película porque me recordó a mi abuelo* contiene un grado de razonamiento y por tanto de información útil, que podríamos traducir, por ejemplo, como *la película provoca demasiada empatía en una parte del público* frente a un comentario de tipo irracional como *la película me pareció horrible*, donde la verdadera razón e información sobre el porqué de dicho sentimiento no está presente en el comentario, si es que dichas razones son realmente conocidas por el usuario.

Como se ha mencionado, el hecho de que un comentario presente indicios de razonamiento, no quiere decir que tal razonamiento sea correcto, pero en sí constituye una fuente de información de cómo un usuario justifica una posición.

Podemos encontrar razonamientos que utilizan lógicas mal construídas, como es el caso de las falacias. Por tanto, si quisiéramos conocer qué comentarios contienen realmente información no solamente útil sino susceptible de ser un *hecho*, deberíamos ejecutar en dos pasos: primero la selección de aquellos comentarios que contengan indicios de racionalidad y posteriormente minar los mecanismos de dicha racionalidad, es decir, identificar primero la *información útil* y a partir de ahí, la *información factual*.

Entendemos que la manifestación de aquello que llamamos racionalidad en los comentarios se produce principalmente en forma de argumentación a través de diferentes sentencias, unas de carácter expositivo y otras de apoyo a las primeras, también conocidas como *evidencias*, conectadas por mecanismos lógicos del lenguaje. En el ejemplo anterior, *No me gustó la película* sería la componente expositiva, *me recordó a mi abuelo* sería la evidencia que apoya la exposición y ambas están conectadas por el conector *porque*. Sin embargo, no todos los casos son tan simples ni todos los mecanismos de argumentación son tan explícitos, ni encontramos los mismos mecanismos de argumentación en todos los contextos. El estudio de la argu-

mentación en textos, cómo identificarla, cómo conectar unos argumentos con otros, etc. constituye un campo de investigación en sí mismo, donde el uso de *corpora* es clave para estas tareas, que también nos servirán de utilidad en nuestro trabajo.

De entre todos los temas sobre los que los usuarios pueden opinar, existe un conjunto de temas identificados como *controvertidos*, que son de especial interés a la hora de extraer una información útil y factual sobre el tema en concreto.

Entendemos como *controversia* al debate público y reiterado entre diferentes grupos de individuos que defienden posiciones contrarias, usando diferentes argumentos para defender sus posiciones. Entre dichos temas controvertidos encontramos algunos como el aborto, la pena de muerte, el cambio climático, el uso de algunas terapias o productos, la contaminación, etc.

Por supuesto, la controversia tiene un aspecto contextual sociológico e histórico en muchos temas. Mientras que la pena de muerte es aún un tema controvertido y de debate en Estados Unidos, no es así en España donde es un tema considerado como cerrado, donde se ha alcanzado un *consenso*: la mayoría de la comunidad considera válidos ciertos argumentos (a modo de verdad) y aunque una parte de la población aún defienda una posición contraria, su condición de minoría no es suficiente para mantener viva dicha *controversia*.

Algo que tienen en común todos los temas controvertidos es que, *a priori*, no existe una relación, justificación o causalidad suficiente para concluir que existe una verdad *de facto* sobre el tema. Ya sea porque no tenemos aún suficiente información sobre el tema, o pruebas que la comunidad considere concluyentes o bien porque en tal tema no es posible llegar a un punto de consenso. Por ejemplo, en cierto contexto, podría darse que el tema de *la existencia de Dios* pueda considerarse controvertido. Existirían dos grandes posiciones enfrentadas, los que están a favor y en contra de esa existencia, pero no existen mecanismos lógicos que nos hagan llegar a un consenso. Aunque para ese contexto concreto, podría aceptarse finalmente una de las posiciones como verdad a medida que el debate avanza y dicha posición actuase entonces como una *verdad de facto*.

La detección de controversia en la Web puede resultar una herramienta muy útil como una medida de calidad de la información y como una herramienta más para el usuario, si es informado de que un contenido tiene cierto grado de controversia, puede desarrollar una visión más crítica y menos sesgada de la información que consume y por tanto ayudar a prevenir los fenómenos de sesgo de información, dirigida o espontánea, mencionados antes.

Si un tema aparece como controvertido en una comunidad de individuos, puede ser un indicador de que la información disponible respecto a ese tema aparezca como sesgada o fraudulenta ante una comunidad dividida en diferentes grupos de opinión, que resulta polarizada ante esta información. Esto puede darse bien cuando los diferentes grupos de opinión de la comunidad se encuentran en posesión de diferente conocimiento, tienen diferente criterio ante una misma información o habitan en

contextos diferentes. Pero también puede darse cuando el tema de controversia no está bien definido, y en algún sentido esto manifiesta sesgo de información.

Existen, como ya hemos visto, diversos ámbitos en los que la controversia, el sesgo y la información sesgada están presentes: religión, política y geopolítica, deportes, ciencia e investigación, salud, etc. De entre ellos, podríamos considerar algunos casos que son verdaderamente críticos para la sociedad, como es el caso de la salud pública.

En el segundo caso, recomendaciones basadas en información sesgada, pseudociencias, campañas de marketing sobre fármacos, efecto placebo y las dificultades para demostrar en algunos casos la verdadera eficacia de un fármaco, de una terapia, de un hábito o de un ingrediente milagroso, pueden causar problemas de salud y verdaderas crisis sanitarias en poblaciones que no aplican criterios estrictos y análisis crítico sobre la información que consumen.

El ámbito médico en general, ha sido un ámbito plagado de controversia mucho antes del uso de Internet. Probablemente debido a la gran carga de superstición, costumbres, ideología, ética y otros aspectos presentes en la sociedad.

La consolidación tardía de la Medicina como campo formal en muchas sociedades, junto a la falta de información que aún se tiene sobre muchos aspectos de nuestro organismo, además de la componente sociológica y psicológica de las enfermedades y los síntomas, hace que el grado de controversia esté mucho más presente que, por ejemplo, en el campo de la biología u otras ciencias formales.

Particularmente nos interesa el tema del uso de fármacos, ya que, aunque estos pasen internamente su proceso de pruebas necesarias para poder salir al mercado, éstas se hacen sobre muestras de la población que deberían representar significativamente a la población real. Dicha representatividad es realmente complicada de obtener para el caso de ensayo clínico, por la gran cantidad de variables que están en juego, aspectos ambientales, genéticos, de compatibilidad con otros fármacos y otros estados de salud, de alimentación, etc.

Adicionalmente en la actualidad, a pesar de los medios existentes, no existe un método estandarizado de recopilar *feedback* sobre el uso de fármacos de manera consolidada. Podemos encontrar diferentes foros especializados, redes sociales e incluso contacto directo de los usuarios con la industria; pero sería de gran utilidad el desarrollo de herramientas que puedan detectar en diversas plataformas de la Web, las diferentes experiencias de los usuarios respecto al uso de fármacos y ayudar así a entender mejor el uso final e identificar aquellos fármacos que no cumplan con la función para los que son prescritos.

La detección de controversia en la actividad de los usuarios en la Web, a través de comentarios y opiniones en diferentes plataformas sobre el uso de fármacos, puede constituir una herramienta de gran utilidad, inexistente hasta el momento, que nos puede dar una mejor idea de la interacción de los usuarios con sus fármacos a través de sus experiencias y constituir un primer paso para extraer una información más

objetiva y factual sobre el uso de dichos fármacos y sus efectos.

De esta manera, podría como ya se ha mencionado, ayudar a combatir la información sesgada en la Web y dar al usuario una herramienta extra para enriquecer su juicio a la hora de evaluar la credibilidad y objetividad de la información que consume, particularmente en el ámbito farmacológico pero potencialmente también en otros ámbitos.

Este es el caso que precisamente hemos considerado en este trabajo, por su interés para el bien público, ya que la desinformación en el campo de la salud puede tener consecuencias verdaderamente devastadoras y cada vez son necesarias nuevas técnicas que nos permitan explotar todos los aspectos de la información disponible para combatir dicha desinformación, en este caso mediante técnicas de procesamiento de lenguaje natural; por la disponibilidad de corpora para desarrollar el trabajo en este ámbito y por su potencial extrapolación, desarrollo y extensión en el futuro; por la actualidad de la problemática del sesgo de información y por la actualidad de la industria farmacéutica en los últimos tiempos (y meses).

2.2. Detección de Controversia

Una primera investigación que afronta el problema de indentificar temas controvertidos de manera automatizada en textos es la de (Choi, Jung, y Myaeng, 2010), concretamente en artículos de noticias.

El primer paso para desarrollar una aproximación al problema es, efectivamente, desarrollar una serie de definiciones necesarias para consolidar el marco teórico. El primer concepto a definir es el de *tema controvertido*, definido como un *concepto que invoca sentimientos o visiones en conflicto para dicho tema*. Para garantizar la coherencia de lo que se entiende por tema, es definido toda pieza de texto que puede obtenerse mediante una *query* en un motor de búsqueda. A pesar de que podría ser cualquier texto de longitud arbitraria, en el estudio se limita a considerar segmentos nominales o verbales, como puede ser, por ejemplo, *la guerra de Afganistán*.

El método para realizar dicha identificación consiste, por tanto, en medir la magnitud del sentimiento presente y la diferencia entre las cantidades de dos polaridades diferentes. Adicionalmente, en un segundo paso, se pretende identificar subtemas, definidos como entidades o conceptos que están conceptualmente asociados y subordinados al tema de controversia y sirven de pilar para apoyar las posiciones adoptadas por las diferentes polaridades.

De aquí se deriva el primer aspecto de la controversia que tendremos en cuenta, la existencia y confrontación entre facciones, caracterizada por una polaridad determinada. Sin embargo, para (Choi, Jung, y Myaeng, 2010) se trata del único aspecto de la controversia a tener en cuenta. Se asume que un tema controvertido provoca polaridad (positivo vs. negativo, moral vs. inmoral, correcto vs. incorrecto, etc.) y que existen unos subtemas, representados por segmentos de texto con carga

sentimental, que sirven de razones para apoyar una de las posiciones.

Lo que hace único el método presentado en (Choi, Jung, y Myaeng, 2010) es que no se trata meramente de identificar sentimiento en bloques temáticos de texto, sino de identificar temas que son evocadores de sentimiento y sus sub-temas, lo cual se identifica directamente como semillas de controversia.

Para un documento dado, a través del método presentado en (Azzopardi, De Rijke, y Balog, 2007) de generación de queries, se extraen términos que son evaluados en sentimiento mediante el score dado por *SentiWordNet*. Mediante la comparación de las magnitudes dadas para las polaridades positiva y negativa de un conjunto de términos, utilizando una parametrización experimental a modo de umbral para decidir cuando el conjunto de términos dado es suficientemente controvertido o no. Esta comprobación viene a raíz del corolario evidente de la definición de controversia dada, mediante el cual, un tema no será controvertido si la polaridad generada es ampliamente positiva o negativa. Situación que encontramos si consideramos, por ejemplo, un tema como *la gripe del cerdo*.

De estos puntos, podemos destacar la importancia de la selección de las unidades de texto para evaluación y la utilidad de comparar polaridades de sentimiento para extraer un grado de enfrentamiento para dicho tema. Ambos aspectos serán anallizados y tenidos en cuenta en nuestra investigación.

Para la extracción de subtemas, se realiza un *parseado* de los textos, extrayendo segmentos nominales que son clasificados a través de un clasificador estadístico, basado en un conjunto de características, donde algunas de ellas son básicas, como si los términos aparecen en el título o si podrían ser el sujeto u objeto de un título de otro artículo, y otras estadísticas como medidas de similitud al tema y al contexto.

Respecto a los datos utilizados en los experimentos, se ha utilizado el corpus MPQA, que contiene artículos de noticias de 187 fuentes, considerando los años 2001 y 2002, consistente en 355 documentos (8955 oraciones, marcadas como polarizadas o no polarizadas), clasificados en 10 temas diferentes, que son considerados *a priori* como controvertidos.

Para la evaluación de la extracción de subtemas se ha extraído y anotado manualmente un subset del corpus, mediante tres anotadores y resolución de conflictos mediante voto mayoritario.

Como trabajo análogo y casi simultáneamente, encontramos el estudio (Popescu y Pennacchiotti, 2010), centrado en detectar eventos controvertidos en Twitter. Se trata del primer trabajo en el ámbito de la detección de controversia que aplica al caso de una red social y contenido generado por usuarios, a diferencia del contenido tratado en el artículo anterior. En este caso, la controversia se analiza a través de los llamados *eventos controvertidos*, que son considerados un tipo de evento interesante (*engaging*) para los usuarios. Se entiende por *evento controvertido* aquellos eventos de *Twitter* que provocan una discusión pública en la que sus miembros expresan opiniones contrarias (en vez de indiferencia o amplio consenso en una dirección u

otra).

Esta definición de la característica de *controvertido* coincide con la presentada anteriormente, aplicada al ámbito de las redes sociales, donde un *evento* consiste, dado una entidad objetivo concreta, en una actividad o acción con una duración clara y finita, con dicha entidad como protagonista.

Se considera también el concepto de *Twitter snapshot*, como el triplete formado por una entidad objetivo (por ejemplo, *Barack Obama*), un periodo de tiempo dado (por ejemplo, 1 día) y un conjunto de *tweets* sobre esa entidad en ese periodo de tiempo. Dentro de un *snapshot* pueden encontrarse tanto eventos controvertidos como no-controvertidos (spam, discusión genérica, etc.)

El proceso de detección de eventos controvertidos puede describirse en dos pasos: asignación de un *score* de controversia a cada *snapshot* y realizar un *ranking* de los *snapshots* según su medida de controversia.

La modelización se realiza como una tarea de *Machine Learning* supervisada, donde cada *snapshot* está representado por un vector de *features* construido a partir de *Twitter* y otras fuentes. Encontramos básicamente dos variaciones del método propuesto:

- Modelo directo: estimación del *score* de controversia en un solo paso mediante el uso de un modelo de regresión de *Machine Learning*. El set de entrenamiento estará formado por ejemplos positivos *snapshots con eventos controvertidos*.

- Modelo de *pipeline* en dos pasos: primero se realiza la detección de eventos mediante un modelo de clasificación que selecciona *snapshots* con eventos a partir del set total de *snapshots* y a continuación se evalúa el nivel de controversia de los *snapshots* seleccionados.

Es de gran interés las definiciones que se proponen, entre las que encontramos la primera definición explícita de controversia que identificamos, $Controversy = \frac{Min(|Pos|,|Neg|)}{Max(|Pos|,|Neg|)} \cdot \frac{|Pos|+|Neg|}{|Pos|+|Neg+|Neu|}$, basada en las polaridades positivas, negativas y neutras de conjuntos de *tweets*. Sigue siendo una definición muy dependiente del uso de *Twitter* como fuente de información, pero conceptualmente nos aporta una referencia para realizar posterioremente una definición más amplia.

Una extensa línea de investigación es la desarrollada por Dori-Haconen et al. a través de trabajos como (Dori-Hacohen y Allan, 2013), (Dori-Hacohen y Allan, 2015) o (Dori-Hacohen, Yom-Tov, y Allan, 2015), se reflexiona sobre la necesidad de una definición amplia y técnica de controversia, que efectivamente no hemos comprobado aún, y que sería uno de los primeros pasos de nuestro trabajo, aunque hemos comprobado que la mayoría de las investigaciones parten de una idea generalizada en común: la controversia como un escenario polarizado, enfrentado y sin consenso.

La línea de Dori-Haconen se centra principalmente en la detección de controversia en artículos de Wikipedia y su potencial extensión en sistemas análogos a Wikipedia. En estos sistemas, el objeto de controversia se ve a menudo representado por un artículo o artículos, que se encuentran ligados a otros artículos mediante sus referencias. Se dispone además, de metadatos sobre dichos artículos, que permiten realizar en principio una estimación de controversia. En el caso de Wikipedia, tenemos por ejemplo el número de ediciones y correcciones cruzadas sobre un mismo artículo, síntoma de la falta de consenso. Se asume además que si un artículo o tema es controvertido, esta controversia se ve extendida hasta cierto punto también a los vecinos que están conectados a él.

Podemos remarcar pues los siguientes puntos a raíz de esta línea de investigación: la falta de una definición ampliamente aceptada y técnicamente aplicable de controversia, la dependencia del ámbito en el que se desarrolla dicho mecanismo de detección y la necesidad de anotadores humanos con unos criterios claros de anotación y evaluación para supervisar el proceso.

Respecto al segundo punto, cabe mencionar que en este caso el proceso entero se basa en las propiedades específicas, estructura y datos aportados por una plataforma como Wikipedia, quedando muy limitadas las posibildades de extrapolación de dichas técnicas a otros ámbitos.

La detección de controversia se define como un problema de clasificación binaria (controvertido vs. no-controvertido), donde el objeto de controversia es el contenido de una página Web (en este caso, un artículo de Wikipedia).

Cabe destacar en (Dori-Hacohen, Yom-Tov, y Allan, 2015) una breve mención al uso de *argument mining* como una vía para aproximar el problema, bajo la premisa de que la confrontación y polaridad descritas se pueden identificar con un escenario de intercambio de argumentos a favor y en contra de una posición o un tema.

Este concepto es desarrollado más ampliamente en (Addawood y Bashir, 2016), donde se propone directamente como método para analizar la controversia en redes sociales. Se destaca la importancia de analizar la estructura interna de la argumentación, la cual se basa en la existencia de dos tipos elementos *claims* y *evidences*, que podríamos describir como la exposición de una idea y los elementos que la apoyan o atacan. A continuación, el análisis se centra en clasificar qué tipo de elemento se presenta como prueba para apoyar o atacar la argumentación (opinión experta, noticias, entrada de blog, fotografía, etc.)

Gracias a un conjunto de *features* léxicas, gramaticales, un diccionario psicométrico y *features* propias de *Twitter* se consiguen resultados del 89,2 % en clasificación binaria (argumento/no-argumento) para un set de 3000 *tweets*.

Sin embargo, tenemos una enorme dependencia del uso de *Twitter*, sus *features* y el tamaño reducido del dataset utilizado, que nos hace dudar que una aproximación de este tipo puede ser generalizada con éxito. A pesar de esto, la aproximación de usar la minería de argumentos como base para el estudio de la controversia

nos resulta interesante, ya que cualitativamente, la confrontación polarizada entre varias partes es un escenario propio del cruce de argumentaciones. De esta manera también *filtraríamos* aquellos elementos textuales que contienen información que puede ser de poca relevancia, o que pueden alterar la polaridad del conjunto sin aportar una posición justificada.

Desde nuestro punto de vista, el uso de una sub-clasificación, o mayor granularidad para la categoría de argumento añade una complejidad que no se ve compensada y que puede introducir ruido en la clasificación final, ya que esta sub-clasificación debe ser también evaluada.

Existen definiciones adicionales y propuestas de algoritmos para detectar la controversia (Bykau et al., 2015), pero de nuevo para el caso *Wikipedia* y para propiedades propias de *Wikipedia*, como el análisis de las *guerras de edición*, en la que la comunidad edita y re-edita secciones de artículos que son considerados controvertidos.

También en el ámbito de Wikipedia, podemos encontrar métodos que explotan más concretamente la componente sentimental, ya mencionada, en el contenido del texto. Es el caso de (Wang y Cardie, 2016) donde se propone una detección de disputas online a través de identificar secuencias de oraciones cargadas de componente sentimental que se usan en un clasificador que asigna la etiqueta de *disputa / no-disputa* para la discusión como un todo. Adicionalmente de las *features* de sentimiento, se usan *features* típicas de los dominios del dataset, *features* léxicas y *features* típicas de los mecanismos de discusión (número de turnos, participantes, número medio de palabras usadas, número de revisiones, etc.)

Conscientes de la complejidad de identificar la controversia debido a la variedad de formas en las que aparece y la variedad de los ámbitos en los que se puede encontrar, aparte del ya mencionado y tratado caso de Wikipedia, se propone en (Linmans, van de Velde, y Kanoulas, 2018) un enfoque semántico basado el uso de *word-embeddings* (Mikolov et al., 2013) para entrenar modelos de redes neuronales que puedan captar las sutilezas semánticas de la controversia.

Los resultados consiguen mejorar las aproximaciones basadas en modelos léxicos, sobre todo al comparar los resultados obtenidos de dos periodos de tiempo diferentes, demostrando una mayor capacidad de adaptabilidad frente a cambios en los datos.

El uso de *embeddings* en este problema es delicado, ya que a pesar de que estamos capturando el aspecto semántico del texto considerado, la controversia puede tener un grado mayor de sutileza, en el que el sentido de las palabras puede ser clave y este no es capturado por el *embedding*. La dificultad de capturar aspectos de múltiples contextos posibles a través de un *embedding* y la falta de corpora de carácter amplio anotados para el problema de la controversia añaden complejidad a este enfoque.

Finalmente, se han desarrollado diversos trabajos sobre la cuantificación de

la controversia como (Garimella et al., 2018), que además se realizó en el ámbito de las redes sociales, al igual que en nuestro caso. Como hemos podido observar, normalmente encontramos aproximaciones en las que se pretende identificar de manera prácticamente binaria la presencia de controversia en un contenido y la minería de este contenido en sí para un mejor entendimiento de las razones de la controversia. Sin embargo, la cuantificación misma de la controversia ha quedado en segundo plano prácticamente hasta (Garimella et al., 2018), siendo superficialmente asociado simplemente a una medida de de la polarización de los individuos hacia el tema de controversia.

La cuantificación de la controversia resulta un paso necesario también para nuestro trabajo, ya que entendemos que la controversia no se mostrará siempre de manera clara en forma binaria (controvertido vs. no-controvertido), sino de manera gradual y cuantificable según nuestra definición y que esta puede ser posteriormente traducida a un paradigma binario mediante un mecanismo de *thresholding*.

Uno de los objetivos de (Garimella et al., 2018) es el de desarrollar una metodología independiente del ámbito, a diferencia de otras muchas de las líneas comentadas anteriormente, fuertemente ligadas a un ámbito y forma del contenido. El método está basado en tres pasos donde se aplica el uso de grafos: construcción de un grafo de la conversación sobre el tema; particionado del grafo de conversación para identificar las facciones potenciales de la controversia; y medida de la cantidad de controversia a partir de las características del grafo, mostrándose que las características de contenido no son tan eficientes para capturar la controversia como lo es esta aproximación.

La hipótesis de la que parten es que es posible realizar este análisis del grafo de la conversación y detectar una estructura de *clusters* para un tema determindo. Esta hipótesis se basa en el hecho de que un tema controvertido implica diferentes partes con puntos de vista opuestos y que los individuos del mismo bando tienden a amplificar y apoyar sus argumentos entre sí.

Para ello, de nuevo basándose en la aproximación de grafos, se extraen diversas *features* basadas en aspectos como: la estructura de los apoyos, que describe, por ejemplo, quién apoya a quién; la estructura de la red social, que describe quién está conectado con quién en la conversación; el contenido descrito a través de palabras clave usadas en el tema; el sentimiento, es decir el tono positivo o negativo para discutir el tema. De entre todas las *features* estudiadas, aquellas relacionadas con el contenido parecen ser las menos relevantes. A partir de estas *features* se realiza el cómputo del *score* de controversia para el tema, donde aquellas derivadas de las propiedades de grafo y aquellas relacionadas con el sentimiento parecen ser las más relevantes.

En definitiva, el trabajo realizado por (Garimella et al., 2018) se aproxima bastante a lo que queremos desarrollar en este trabajo. Sin embargo, a pesar de ser un trabajo que profundiza mucho más en un método y concepto para la identificación

y cuantificación de la controversia que resulte independiente del tema, corpora y ámbito, aún sigue siendo un enfoque muy ligado a un escenario de una red social como es *Twitter*.

El enfoque llevado a cabo a través de elementos de la teoría de grafos es especialmente adecuado para abarcar el escenario de redes sociales. Pero en este trabajo, se pretende acercarnos a una definición y metodología amplia para la identificación y cuantificación de controversia, partiendo de un caso de estudio, pero sin perder de vista el enfoque generalista.

La definición de controversia que adoptaremos es en general prácticamente idéntica a la descrita en los trabajos mencionados: *un escenario de discusión pública sobre un tema particular en el que varios grupos de opinión, posición o polaridad toman parte y contrastan sus visiones en contraposición.* Sin embargo, debemos concretar un poco más qué significan estos términos a la hora de desarrollar un pipeline. En nuestro caso, consideramos que la base de una discusión sobre un tema controvertido debe ser la argumentación, no solamente como medio para extraer las diferentes razones que alimentan las posiciones enfrentadas, sino también como condición necesaria pero no suficiente para que se dé el fenómeno de la controversia. Ampliaremos esta definición más tarde, pero, de momento, podemos partir de la idea de que la controversia es un fenómeno que se da cuando existe un razonamiento detrás de las posturas adoptadas por las partes. Podemos así, empezar a distinguir entre un escenario de controversia hacia un tema y simplemente un escenario de desacuerdo cargado de elementos no-racionales y/o sentimentales. Descartamos así escenarios como aquellos en los que los usuarios de una red social interaccionan entre sí, incluyendo réplicas, apoyos y polaridad, pero su contenido no contiene indicios de razonamiento o argumentación. Adicionalmente, en contraste con el trabajo de (Garimella et al., 2018), nuestro enfoque no está basado en métodos de la teoría de grafos y en ese sentido tiene mayor capacidad de generalización a otros casos, ya que la asunción de la que parte el enfoque basado en grafos es que existe, por un lado, una red social representable mediante nodos (usuarios) que están conectados entre sí formado una topología y, por otro lado, el grafo que representa la estructura conversacional del debate (quién responde a quién). A pesar de los buenos resultados obtenidos por (Garimella et al., 2018), el enfoque pierde parte de su aplicabilidad si no existe una estructura de conversación definida o una conectividad explícita entre los usuarios.

En el caso de estudio considerado para este trabajo se considera un conjunto de comentarios de usuarios hacia un producto, sin conocer las características de la plataforma en la que los usuarios interactúan (en nuestro caso sólo conocemos que provienen de foros del ámbito médico) ni la posible relación existente entre usuarios. Cada comentario consistirá en una unidad independiente a considerar, con la única característica de haber sido realizada por un usuario respecto a un tema concreto (en este caso, el uso de un fármaco para un caso de uso concreto).

En este escenario descrito, puede darse una situación de controversia sin que los usuarios tengan que entablar realmente una interacción en forma de conversación, sino que estarían apuntando sus interacciones hacia el objeto de la controversia en sí. No obstante, sí que existiría una estructura conceptual de argumentos que interaccionan entre sí donde la teoría de grafos podría tener un gran papel.

Para nuestro propósito, realizaremos un análisis básico, a través de un algoritmo de *clustering* y una posterior evaluación de sentimiento de dichos clusters. Nuestro objetivo es el de proponer y evaluar un *pipeline* lo suficientemente general y adaptable, sin grandes dependencias de conocimiento previo de la plataforma, ámbito o contenido que permitan realizar una estimación de la controversia de un tema ante una comunidad de usuarios. Por tanto, la profundización y perfeccionamiento de técnicas que aplican a casos particulares no han formado parte de este desarrollo como elementos principales, sino más bien de elementos complementarios que pueden enriquecer este método de base.

A partir de las características en torno a estos tres elementos mencionados: argumentación, agrupación y polaridad; realizaremos una estimación de controversia basada en nuestra definición teórica y representada como diferentes pasos de nuestro *pipeline*.

Sí que conservamos el concepto de el *clustering*, ya que aunque no conozcamos si existe una conexión entre usuarios, la existencia de grupos de opinión que combinan una componente de contenido y otra componente de polaridad es necesaria para la existencia de controversia, coincidiendo ampliamente con la mayoría de los enfoques considerados en el trabajo previo.

Nos interesan particularmente las capacidades de adaptar este enfoque para otros casos y en escenarios como *Twitter*, la aproximación por grafos puede resultar un complemento perfecto para una mayor granularidad del concepto de controversia. Por ejemplo, como presenta (Garimella et al., 2018), la capacidad de estimar cómo de controvertido es un usuario concreto.

Respecto a las definiciones de controversia, aún compartiendo un concepto general y *humano* de lo que entendemos por controversia, no se ha encontrado una definición lo suficientemente amplia y fácilmente aplicable, dependiendo en la mayoría de los casos del ámbito particular o de las técnicas elegidas. Uno de nuestros objetivos, y primeros pasos, es definir ampliamente el concepto de controversia y derivar a partir de él una medida práctica para su estimación a partir de las propiedades que extraigamos de los datos.

En nuestro caso, como ya se ha mencionado, partimos de la base de que los elementos considerados para estudio deben contener indicios de argumentación, y a partir de ahí desarrollamos, por un lado la agrupación de dichos argumentos que constituirán las diferentes posiciones en confrontación y la detección de su polaridad para cuantificar dicha confrontación.

Para ello, analizaremos el estado del arte de las tareas implicadas en las con-

diciones de nuestra definición: **detección argumentación**, **clustering** y **análisis de sentimiento**.

2.3. Extracción de Argumentos

En muchos de los trabajos analizados en cuanto al *estado del arte* de la detección de controversia, hemos encontrado numerosos intentos de captar las *features* características de la controversia a través de diferentes aproximaciones. De algunas de ellas (Dori-Hacohen, Yom-Tov, y Allan, 2015)(Addawood y Bashir, 2016) se deriva la noción de que el problema de la detección de controversia, contiene en algún sentido un problema de detección y extracción de argumentación, ya que existe un amplio consenso en que la controversia surge de la confrontación de grupos de opinión o de contenido en sí que se manifiesta en uno u otro sentido hacia el tema en cuestión.

Ante la dificultad de definir de manera precisa el problema de la detección de la controversia (e incluso la controversia en sí), podemos considerar el problema de la detección de controversia como un subtipo del problema de detección de argumentos. Este problema, según se ha comprobado en el estado del arte, está mucho más definido, acotado y estudiado que el anterior. A partir de la detección de argumentación, podremos acotar o matizar aún más nuestra aproximación para que satisfaga las condiciones que se manifiestan en el caso de la controversia, como son la polaridad o la existencia de estructuras o grupos. Adicionalmente la argumentación tiene un carácter más amplio y homogéneo *cross-* dominio, ámbito, fuente de información, etc.

Desde nuestro punto de vista, como ya se ha mencionado en la sección anterior, los indicios de argumentación deben estar presentes para filtrar todo el contenido *racional* de aquel contenido *irracional* o meramente de expresión de un sentimiento sin llegar a tener una argumentación o una meta tras de sí.

Como se describe en (Stab y Gurevych, 2014b), las investigaciones en el campo de *Argument Mining* se han centrado en diferentes tareas, dado un contenido textual separado en unidades de texto definidas: la clasificación de unidades de texto en argumentativas vs. no-argumentativas; la definición y separación del objeto argumento en sub-componentes con un esquema definido y la identificación de estructuras de argumentación.

La distinción entre lo que es argumento y no en un texto se considera normalmente un problema de clasificación binaria, como podemos ver en los primeros trabajos sobre el tema, como son (Moens et al., 2007) o (Florou et al., 2013). En (Moens et al., 2007), se propone una aproximación para identificar oraciones argumentativas en el corpus *Araucaria* desarrollado por (Reed y Rowe, 2004), cuyas anotaciones se basan en la teoría de estructuras de la argumentación propuesta por (Walton, 1996) y constituye el primer corpus anotado con este propósito.

Es crucial en estas aproximaciones identificar correctamente cuáles son los *features* más relevantes para extraer e incluir en los consiguientes modelos de clasiicación usados así como un corpus o corpora propiamente anotado para extraer dichas categorías. Los resultados obtenidos tanto por (Florou et al., 2013) como por (Moens et al., 2007) alcanzan en el mejor de los casos un *F1* del 76 % al detectar si un segmento de texto se considera argumentativo o no. Estas features suelen consistir en secuencias de tokens, estadísticas sobre el texto, puntuación, verbos y tiempos verbales, *keywords* y marcadores del discurso.

Respecto a la tarea de identificar los componentes de la argumentación, encontramos una primera aproximación en el trabajo de (Teufel y others, 1999), el cual tiene como objetivo la segmentación de un texto para la construcción posterior de un resumen de su contenido (para el caso concreto de artículos científicos), en el que cada oración se clasifica en uno de siete roles retóricos, los cuales incluyen *premisa*, *resultado* o *propósito*.

A partir de este punto, existe un amplio consenso, sobre la estructura interna de un argumento, consistente principalmente en dos tipos de componentes: *claims* y *evidences*, donde los primeros son afirmaciones que contienen la postura en sí del argumento hacia el tema en cuestión y los segundos son elementos que justifican los primeros (se hablará en detalle en la sección 2.3.1).

Podemos destacar a partir de los trabajos y líneas de investigación analizadas, la existencia de tres puntos clave a la hora de realizar una tarea completa de minería de argumentos:

1. La clasificación de unidades de texto en argumento vs. no-argumento.
2. La identificación de los sub-componentes de la argumentación: principalmente *claim* y *premises* o *evidences*.
3. La identificación de la estructura de la argumentación.

Dado el marco en el que se desarrolla este trabajo, la componente argumentativa necesaria para nuestros objetivos, se reduce prácticamente al primer punto, la detección de argumentación en un texto. A pesar de que en muchos de los enfoques mencionados se profundiza más respecto a si los argumentos encontrados son favorables o contrarios al tema que se trata en el documento, a el hecho de encontrar indicios de argumentación y poder clasificar la unidades de texto como *argumentativas* o *no-argumentativas* puede resultar suficiente para nuestros objetivos.

Bajo esta hipótesis, de que la argumentación puede tratarse como un problema de clasificación, surge la necesidad de definir un conjunto de *features*, un modelo de clasificación y sobre todo un corpus, previamente anotado con las categorías a identificar. Este último punto, aunque muy dependiente de las otras dos, será analizado en la sección 2.3.1.

En (Aker et al., 2017) encontramos una buena visión general y descripción del proceso. Se consideran dos corpora de referencia: (Aharoni et al., 2014) y los ensayos

persuasivos mencionados en (Stab y Gurevych, 2014b), analizados posteriormente; una serie de *features* propuestas por el trabajo de (Stab y Gurevych, 2014b) consistentes en diferentes tipos: estructurales, léxicas, sintácticas, indicadores de la argumentación, contextuales, *word-embeddings*, etc.; se realiza una tarea de clasificación binaria a nivel de sub-componente del argumento (*evidence*, *claim*, *none*), utilizando diferentes modelos de clasificación: Regresión Logística, *Support Vector Machine*, *Random Forest*, *Naïve-Bayes*, *Adaboost*, *Convolutional Neural Networks*, etc. alcanzando unos resultados de $F1 = 66\,\%$ con el para un set de 3832 ensayos persuasivos (Stab y Gurevych, 2014b) y de $F1 = 94\,\%$ para el caso de 2858 instancias de datos de *Wikipedia* (Aharoni et al., 2014). El contraste de los resultados puede interpretarse como una medida de la complejidad de la tarea. En el caso de los ensayos persuasivos, encontraremos un estilo más libre, personal y sin tantas pautas ni estructuras como en el caso de *Wikipedia*. El estudio de este caso, por tanto, será de más interés si lo que queremos es conseguir una metodologías amplia que valga para casuísticas diferentes.

En esta línea, encontramos el enfoque de (Stab et al., 2018), en el que se presenta un corpus de nuestro interés, consistente en 25,000 instancias anotadas a nivel de oración como (*argument-pro*, *argument-against*,*no-argument*) para 8 temas diferentes considerados controvertidos. En este caso, no se ha realizado ingeniería de *features*, sino que se utiliza un *word-embedding* de 300 dimensiones a partir de *word2vec*(Mikolov et al., 2013) para representar vectorialmente las instancias y se utilizan diversas versiones y configuraciones de redes neuronales del tipo Long-Short-Term-Memory (Tai, Socher, y Manning, 2015) consiguiendo como $F1 = 66{,}62\,\%$ en el mejor de los casos para clasificación binaria y un $F1 = 42{,}85\,\%$ para la clasificación descrita de tres categorías.

Encontramos de extrema utilidad este último corpus, ya que se adapta a las condiciones requeridas en nuestro caso. Sin embargo, la metodología propuesta no parece mejorar los resultados, a pesar de la creciente complejidad de los modelos de clasificación utilizados.

2.3.1. Corpora para Extracción de Argumentos

Uno de los puntos claves de la detección de argumentación, además de extraer un conjunto de *features* relevantes para el caso, es la selección de un corpus, que por sus características, contenido, tamaño y tipo de anotación se ajuste mejor al caso. Analizaremos a continuación las características de los tres corpora más mencionados en la sección anterior.

2.3.1.1. Wikipedia

El dataset, desarrollado en IBM por Aharoni et al. (Aharoni et al., 2014) se presenta como una manera nueva y única de enfocar las estructuras argumentativas. Se trata

de un corpus de cientos de artículos extraídos de *Wikipedia*, que tras un proceso de anotación, dan como resultado un conjunto de 2683 elementos argumentativos en 33 diferentes temas considerados controvertidos.

En este caso, encontramos una definición de **argumento** como una estructura compuesta por dos elementos, sin tener estos que constituir oraciones completas, referidos como **claim** y **evidence**, en alusión a su propiedad funcional dentro de la argumentación.

En el modelo propuesto, un **topic** es normalmente una sentencia corta que define el sujeto de interés. Un **Context Dependent Claim** es una sentencia general que apoya o es contraria al *topic*, mientras que un **Context Dependent Evidence (CDC)** es un segumento de texto que directamente apoya al *CDC* en el contexto del *topic*.

Además, ya que un *claim* puede ser apoyado por diferentes tipos de evidencias. Aquí se consideran tres tipos diferentes de *CDE*:

- **Estudio**: resultados de un análisis cuantativo, proporcionado en forma de datos o conclusiones.
- **Experto**: testimonio de una persona o entidad con conocimiento o autoridad en el *topic*.
- **Anecdótico**: una descripción de un evento específico o ejemplos concretos.

El mayor desafío a la hora de realizar el etiquetado de elementos es la naturaleza sutil y subjetiva de conceptos como *claim* y *evidence*. Para ello, se definieron dos conjuntos de criterios, respectivamente, para *CDC* y *CDE* y un equipo de 20 etiquetadores entrenados realizaron el trabajo con una estricta supervisión. La metodología de etiquetado se realiza en dos pasos: primero, un equipo de cinco etiquetadores realiza el etiquetado sobre el mismo texto de manera independiente y extrae un primer conjunto de candidatos para *CDC* y *CDE*; a continuación otro equipo de otros cinco etiquetadores trabaja independientemente cruzando las listas de candidatos para ambas categorías. Los candidatos confirmados por al menos tres etiquetadores para una categoría son incluidos en el corpus.

Los criterios de etiquetado se desarrollan a partir de los conceptos teóricos iniciales, a través del análisis de ejemplos relevantes. Según estos, dado un *topic*, un fragmento de texto debe ser etiquetado como *CDC* si y sólo si cumple con los siguientes cinco criterios:

- Fuerza: contenido contundente que directamente apoya o desacredita el *topic* dado.
- Generalidad: contenido general que proporciona una idea relativamente amplia.
- Fraseado: está bien construido o necesita al menos un cambio menor para que tenga sentido.

- Mantiene la esencia del texto: mantiene la esencia del texto original del que fue extraído.
- Unidad de *topic*: referencia uno, o como mucho dos *topics* relacionados.

Los criterios para CDE consisten en cuatro criterios análogos a estos cinco, exceptuando el de generalidad ya que obviamente, la *evidence* de un *claim* no puede ser general, si lo apoya concretamente.

El entrenamiento de los etiquetadores es crucial, ya que sin un consenso en una materia tan subjetiva no puede conseguirse una correcta aplicación del criterio.

En definitiva, para 33 temas de debate, un total de 586 artículos de *Wikipedia*, en inglés, se han conseguido 1392 CDCs distribuidas a lo largo de 321 artículos y 1291 CDEs confirmadas: 431 de tipo *Estudio*, 516 de tipo *Experto* y 529 de tipo *Anecdótico*, aunque pueden darse el caso de estar asociadas a más de un tipo.

Algunos de los ejemplos son, por ejemplo, para el *topic*: *The sale of violent videgames to minors should be banned*:

1. (CDC - Pro): Violent video games can increase children's aggression.
2. (CDC - Con): Violdent video games affect children positively.
3. (CDC - Invalid): Violent video games should not be sold to children.
4. (CDE - Estudio): The most recent large scale meta-analysis – examining 130 studies with over 130,000 subjects worldwide – concluded taht exposure to violent.
5. (CDE - Experto): According to some nurse practitioners, stopping substance abuse can reduce the risk of dying early and also reduce some health risks like heart disease, lung disease, and strokes.
6. (CDE - Anecdótico): In April 2000, a 16-year-old teenager murdered his father, mother and sister proclaiming that he was avenging mission for the mainc haracter of the video game Final Fantasy VIII.

2.3.1.2. Persuasive Essays

Desarrollado por (Stab y Gurevych, 2017), presenta una nueva aproximacion para modelar argumentos, sus componentes y relaciones en el caso de ensayos persuasivos en inglés. El esquema de anotación estará basado en anotar separadamente *claims* y *premises*, así como relaciones de apoyo o descrédito.

Nos encontramos con un tipo de texto, el ensayo persuasivo, en el que el autor intenta convencernos de la validez o invalidez de un tema. De nuevo, los mecanismos que caracterizan a este comportamiento son muy sutiles y debemos identificar sus características, componentes y relaciones.

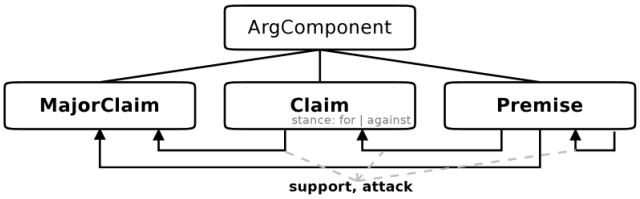

Figura 2.2: Esquema de anotación de argumentos, sus componentes y relaciones

Se entiende por argumento una estructura textual de diferentes componentes, generalmente *claims* (exposición de una idea) y *premises* (elementos que apoyan o desacreditan un claim). Por tanto, a partir de esta definición, podemos entender el proceso de reconocer argumentos en textos como varias subtareas:

1. Separar unidades de texto argumentativas de las no argumentativas.
2. Identificar *claims* y *premises*.
3. Identificar relaciones entre los componentes de la argumentación.

Lo novedoso de este corpus es que intenta profundizar en el caso concreto de la persuasión en textos.

Estos textos persuasivos han sido motivo de estudio previamente en otros contextos, como la *Evaluación Automática de Ensayos*, una línea de investigación que pretende asignar automáticamente una calificación a ensayos de estudiantes por medio de ciertas características.

Uno de esos aspectos es la presencia de argumentación en el discurso, ya que no puede darse una persuasión sin aportar argumentos al respecto. En trabajos anteriores, se ha conseguido delimitar en el texto dónde se pueden encontrar estructuras argumentativas, pero sin llegar a este nivel, de distinguir los diferentes componentes ni las relaciones que se dan entre ellos.

Adicionalmente, se han realizado otros estudios en ensayos persuasivos, centrándose en identificar criterios de estilo, información de hechos, scoring para calidad de argumentación y metáfora. Pero, de nuevo, en ninguno se ha estudiado simultáneamente las subestructuras de la argumentación y ha tenido ensayos persuasivos como objeto de estudio.

Respecto al esquema propuesto, su objetivo es modelar componentes de argumentos así como las relaciones de argumentos, de modo que se propone el esquema mostrado en la figura 2.2.

El estudio de ensayos persuasivos revela una estructura común: normalmente la introducción incluye una *major claim* que expresa la posición general del autor respecto al tema. Esta *major claim* es apoyada o confrontada por los argumentos,

cubriendo ciertos aspectos en los siguientes párrafos. Algunos ejemplos de *major claim* pueden ser (marcado en negrita):

- I believe that **we should attach more importance to cooperation during education**.
- From my viewpoint, **people should perceive the value of museums in enhancing their own knowledlge**.
- Whatever the definition is, **camping is an experience that should be tried by everyone.**

Vemos que existen algunos indicadores de introducción a la *major claim*, en la muchos casos no está tan claro y se requiere de los anotadores que tomen la decisión de anotar la expresión que mejor exprese la posición general del autor.

Los párrafos situados entre la introducción y la conclusión se esperan que contengan los argumentos reales que apoyen o ataquen la *major claim*.

En los siguientes ejemplos podemos ver dos argumentos tipo, conteniendo un *claim* (en negrita) y una *premise* (azul).

- **It is more convenient to learn about historical or art items online.** With Internet, people do not need to travel long distance to have a real look at a painting or a sculpture, which probably takes a lot of time an travel fee.
- **Locker checks should be made mandatory and done frequently** because they assure security in schools, makes students healthy, and will make students obey school policies.

En estos casos, ambas componentes cubren una oración completa, pero esto no es ninguna condición. La anotación se realiza sin tener en cuenta limitaciones, en torno a una sentencia, es decir, una serie de palabras con estructura gramaticalmente correcta. Para indicar si un argumento apoya o ataca la *major claim*, se le asigna un atributo extra con el valor *for* o *against*. De los dos ejemplos anteriores, el primero de ellos se considera *against* por dichos motivos.

En cuanto a las relaciones entre componentes de la argumentación, se distinguen solamente dos relaciones: *support* y *attack*. Ambas relaciones pueden darse bien entre dos *premises*, una *premise* y un *(major-) claim* o entre un *claim* y un *major claim*.

Podemos visualizar esto a través del siguiente ejemplo, representado en la figura Fig.2.3:

- **Living and studying overseas is an irreplaceable experience when it comes to learn standing on your own feet**. One who is living overseas will of course struggle with loneliness, living away from family and friends$_1$ but those difficulties will turn into valuable experiences in the following steps

of life$_2$. Moreover, the one will learn living without depending on anyone else $_3$

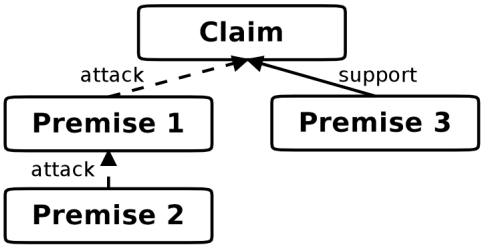

Figura 2.3: Diagrama de la estructura del ejemplo.

Podemos observar cómo el *claim* es atacado por la *premise* $_1$, mientras que *premise* $_2$ es una refutación de la *premise* $_1$. A su vez, *premise* $_3$ apoya directamente el *claim*.

En definitiva, el corpus consiste de 90 ensayos persuasivos, seleccionados de *essayforum*, un foro con una comunidad activa que aporta feedback por escrito a diferentes tipos de textos. Por ejemplo, estudiantes pueden publicar un texto y pedir feedback para evaluar sus habilidades de escritura. Se han seleccionado los ensayos de manera aleatoria y revisado uno por uno, de manera manual reemplazando algunos de ellos cuando no cumplían los requisitos de lenguaje o de contener suficientes elementos argumentativos.

El corpus final consiste de 1673 oraciones con 34,917 tokens. De media cada ensayo tiene unas 19 oraciones y 388 tokens. Contiene exactamente 90 *major claims* (una por ensayo), 429 *claims* y 1,033 *premises*. La proporción entre ambas confirma que en la argumentación un *claim* está apoyado por varias *premises*, como establecían los resultados de Mochales-Palau y Moens (Palau y Moens, 2009), para cubrir varios aspectos y generar una argumentación más sólida.

2.3.1.3. Heterogeneus Controversial Sources

También desarrollado por Stab, Gurevych et al. (Stab et al., 2018), en este caso encontramos un dataset basado en una nueva metodología de anotación de oración cuyo objetivo es poder aplicarse a diferentes dominios y así contribuir a minimizar la dependencia de dominio que encontramos en el campo de la extracción de argumentos.

En este artículo, se aplica *Argument Mining* a la tarea de búsqueda de argumentos. Es decir, buscar argumentos relevantes a un cierto tema dentro de una amplia colección de documentos. Identificar y clasificar argumentos relevantes juega un papel muy relevante en otras áreas como toma de decisiones, razonamiento legal, lectura y escritura crítica y resumen de textos persuasivos.

La automatización de la tarea de búsqueda de argumentos, minimizaría enormamente el trabajo manual que se requiere actualmente. Esta funcionalidad en sí, no ha sido muy desarrollada por ninguna línea de investigación, en parte, por la enorme dependencia de las diferentes aproximaciones y metodologías al ámbito, temática, lenguaje y forma concretos de los textos utilizados. Se consiguen resultados muy buenos en terrenos determinados, pero la aplicación a otros dominios sigue siendo muy deficiente.

Para poder afrontar estos desafíos, en este artículo se propone:

1. Un nuevo esquema aplicable a la perspectiva de búsqueda de información en la búsqueda de argumentos. Se demuestra que es suficientemente generalista para poder ser usado en una gran heterogeneidad de fuentes de datos, y suficientemente simple para ser aplicado manualmente por anotadores sin entrenamiento.

2. Un nuevo corpus heterogéneo anotado, constituido por más de 25000 instancias cubriendo ocho temas controvertidos. Se trata del primer recurso conocido que puede ser usado para evaluar el rendimiento de métodos de Argument Mining cross-temática con fuentes heterogéneas.

3. Una exploración de diferentes aproximaciones para incorporar información en redes neuronales y cómo pueden mejorarse las tareas de aprendizaje cross-topic.

Respecto a otros trabajos anteriores, en muchos de ellos el objetivo ha sido identificar estructuras de argumentación en textos de características y temáticas específicas, como textos de Wikipedia; pero su eficacia frente a otros tipos de texto y dominios es dudosa, y en cualquier caso requeriría de grandes esfuerzos de adaptación.

Se han desarrollado experimentos de Argument Mining cross-dominio solamente para tareas a nivel de discurso tales como identificación de *claims* y identificación de segmentos argumentativos. Pero ninguno que permitiese el uso generalizado sobre diferentes tipos de textos como en este caso.

De nuevo, esclarecemos el concepto de *argumento*, en este caso como un fragmento de texto que expresando evidencia o razonamiento puede usarse para apoyar u oponerse a un tema dado. Un argumento no tiene por qué ser autocontenido, ya que podría estar asumiendo información de un contexto, pero no debe ser ambiguo en cuanto a su orientación hacia el tema.

Un *topic*, por otra parte, es un objeto de controversia hacia el que es posible una polaridad obvia, que suele dar resultados de *a favor* o *en contra* en instancias que se refieren a él.

En algunos modelos de argumentación, como el visto anteriormente, a lo que llamamos *topic* podría ser parte de un *(major) claim* expresando una sentencia positiva o negativa y nuestros argumentos serían a estructura representada por

las relaciones de elementos *premises* apoyando o atacando el *claim* expuesto. Sin embargo, en este caso, a diferencia de las previas representaciones más profundas y complejas a nivel de discurso, este es un modleo plano que considera un argumento como un elemento individual diferenciable del contexto que le rodea.

La primera ventaja es que se reduce enormemente la complejidad de la anotación y de la información y entranamiento necesario para los anotadores.

Para ello, se consideran solamente *topics* que pueden expresarse de manera concisa a través de palabras clave y aquellos argumentos que pueden ser expresados en oraciones individuales. Podemos apreciar estos criterios en los ejemplos de la Fig.2.4, donde es interesante destacar aquellos que han sido clasificados como no-argumentativos, ya que no es suficiente que se posicionen a favor o en contra, sino que también deben contener razonamientos o pruebas en las que apoyar su posición.

Los datos utilizados en los experimentos, han sido recopilados manualmente anotados que cubren una amplia variedad de temas y tipos de textos. A partir de diferentes listas online de teas controvertidos, se han extraido ocho temas. Para cada uno se han realizado búsquedas en Google, se han retirado resultados no archivados en Wayback Machine y se ha truncado la lista de resultados hasta 50.

El resultado es un dataset de documentos Web de contenido polémico incluyendo documentos de noticiarios, editoriales, blogs, foros de debate y artículos de enciclopedia. Se ha realizado un preprocesamiento estándar del texto y además se han retirado todas las oraciones que tengan menos de cuatro tokens o que no tengan verbos. Finalmente queda un de 27,520 oraciones. Las anotaciones se han realizado, clasificando cada oración como *pro-argumento*, *contra-argumento* o *no-argumental* (en las que también se han incluido aquellas demasiado ambiguas o incomprensibles, que solo constituyen menos del 1 %). El procedimiento de anotación se ha ejecutado por dos equipos, uno de expertos anotadores y otros de anotadores sin entrenamiento previo, para clasificar el mismo set de oraciones y así reducir el sesgo. De esta manera, cada oración ha sido anotada independientemente por dos expertos y diez inexpertos. Estas anotaciones son comparadas y si el nivel de acuerdo entre anotadores es superior a 0,721 para la κ de Cohen (Carletta, 1996), la anotación se considera sólida y puede así crearse un *gold-standard*.

topic	sentence	label
nuclear energy	Nuclear fission is the process that is used in nuclear reactors to produce high amount of energy using element called uranium.	non-argument
nuclear energy	It has been determined that the amount of greenhouse gases have decreased by almost half because of the prevalence in the utilization of nuclear power.	supporting argument
minimum wage	A 2014 study [. . .] found that minimum wage workers are more likely to report poor health, suffer from chronic diseases, and be unable to afford balanced meals.	opposing argument
minimum wage	We should abolish all Federal wage standards and allow states and localities to set their own minimums.	non-argument

Figura 2.4: Ejemplos de anotación para el corpus, a nivel de oración para diferentes dominios.

2.4. Clustering de Argumentos

Clustering, también llamado *cluster analysis*, es la tarea de agrupar un conjunto de elementos de manera que los elementos que están en un mismo grupo son más similares entre sí (respecto a algún criterio concreto) que respecto a elementos fuera de ese grupo (o clúster). Es una tarea común del análisis estadístico de datos, utilizada en multitud de casos de uso.

Existe una amplia variedad de algoritmos de *clustering*, que de una manera u otra, se aproximan al problema mediante el cálculo de una medida (o distancia) entre elementos, para a continuación, realizar un proceso de optimización para alcanzar una configuración de grupos óptima. Se trata normalmente de un proceso iterativo, de aprendizaje no supervisado. Es decir, buscamos encontrar relaciones entre ciertas variables, pero no existe realmente una variable objetivo.

En el campo del análisis de texto, se aproxima el problema de *clustering* de documentos o de elementos textuales basándose en diferentes representaciones o *features* de los textos para a continuación aplicar algunos de los algoritmos usuales de *clustering* utilizando dichas *features* o representaciones.

Entre los tipos de algoritmos de *clustering* podemos distinguir dos grandes categorias: los partitivos y los aglomerativos.

En el caso de los algoritmos partitivos, se pretende segmentar el conjunto de documentos en un número predefinido de clusters. Es el caso de algoritmos como K-Means (MacQueen, 1967), K-Medoids (Kaufman y Rousseeuw, 1987) o CLARA (Rousseeuw y Kaufman, 1990). Consideramos principalmente el caso representativo de K-Means, ampliamente utilizado en contraste con otros algoritmos de este tipo.

El algoritmo K-Means recibe como input un número predefinido de *clusters* a identificar, que en algunos casos puede ser previamente estimado de manera heurística o conocido de antemano, y un conjunto de documentos a agrupar.

El proceso consiste en: *(i)* inicializar k elementos del conjunto como centroides provisionales de los *clusters* (existen diversas técnicas de inicialización entre las que se encuentra la generación de k medias aleatorias); *(ii)* asignación de los elementos a los *clusters* por la distancia (euclídea) de cada elemento a la media más cercana; *(iii)* se recalcula el centroide de los k clusters; *(iv)* repetir los pasos *(ii)* y *(iii)* hasta que se dé convergencia.

K-Means es muy eficiente, pero tiene algunos inconvenientes: el uso de este algoritmo implica poder estimar correctamente, o conocer el número de *clusters* que se esperan obtener *a priori*, lo cual no siempre es posible; la convergencia no siempre se alcanza y ésta depende del paso de inicialización.

Las *features* que se suelen utilizar para representar los documentos son las frecuencias inversas de ocurrencias de término (tf-idf) (Balabantaray, Sarma, y Jha, 2015), lo cual no contiene ninguna información semántica sobre los documentos. Este punto junto al hecho d e que no podemos conocer el número de *clusters a*

priori son uno inconvenientes del uso de *K-Means*.

Respecto a los algoritmos algomerativos, han sido igualmente muy estudiados en la literatura (Day y Edelsbrunner, 1984),(Zhao y Karypis, 2002) para todo tipo de datos multidimensionales, entre los que se incluye datos de texto. Es especialmente interesante, ya que va creando una estructura de árbol que aporta información sobre las relaciones y estructura jerárquica de los elementos.

El concepto general sobre el que se basa el clustering aglomerativo es en el de agregar documentos en *clusters* basados en su similitud, normalmente ejecutando sucesivas medidas de similitud por pareja de documentos. La mayor diferencia entre las diferentes modalidades es la manera en la que se calcula dicha similitud.

A medida que se lleva a cabo el proceso en diferentes niveles, veremos cómo se forma una jerarquía de *clusters* (o dendograma) donde encontraremos los documentos iniciales en el nivel más bajo. Cuando dos grupos se fusionan, se crea un nuevo nodo en el árbol, correspondiendo al nuevo grupo formado, y los grupos que lo forman pasan a ser ramas de dicho nodo.

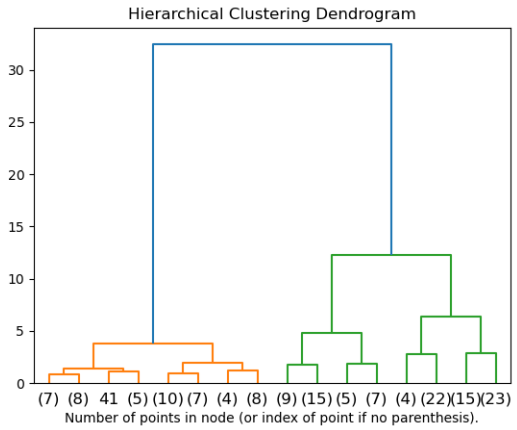

Figura 2.5: Ejemplo de dendograma tras una aplicación de *clustering* aglomerativa jerárquica. (Fuente: https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html)

En (Reimers et al., 2019), podemos ver cómo en el mismo contexto, en la misma línea de *argument mining*, tratada en la sección 2.3, se utiliza para *clustering* de argumentos basada en *clustering* jerárquico aglomerativo, utilizando *word-embeddings* como representación de los argumentos.

Como inciso, los *word-embeddings* fueron introducidos por(Mikolov et al., 2013), en el que se presenta el *toolkit word2vec*, paquete de herramientas que permite manipular y entrenar representaciones vectoriales de textos. A partir de un conjunto de textos, podemos entrenar un modelo de red neuronal para inferir, en base a una palabra, qué palabras deberían formar parte de su contexto o bien, dado un

conjunto de palabras, qué palabras faltan en dicho contexto. De aquí surge una representación vectorial para cada palabra, en un espacio vectorial donde palabras que son semánticamente cercanas tendrán una similitud de coseno elevada.

De esta manera, si tenemos textos anotados con ciertas categorías, podemos representar dichos textos en nuestro espacio vectorial y construir un modelo de aprendizaje supervisado, como hemos descrito anteriormente con modelos tradicionales de *Machine Learning* o bien con otras técnicas de *Deep Learning*.

El uso de estos *embeddings* en nuestro caso será igualmente de utilidad para el *clustering*, ya que nuestro propósito final es identificar los clusters como subtemáticas *de facto*, que comparten un mismo contexto semántico. De manera similar a como se realiza en (Reimers et al., 2019), ejecutaremos experimentos preliminares para detemirnar si una aproximación como *K-Means* puede resultar viable o útil respecto a la comentada aproximación aglomerativa.

2.5. Sentiment Analysis

El análisis de sentimientos es uno de los problemas más estudiados del Procesamiento del Lenguaje Natural, análisis de texto y lingüística computacional. Generalmente, este análisis pretende determinar cuál es la posición del autor de un texto sobre un tema concreto. Para ello, se pretenden clasificar los diferentes documentos, o elementos textuales, en diferentes categorías de sentimiento. Aunque usualmente, se suele utilizar la versión binaria, que reduce considerablemente la complejidad de la clasificación: *positivo* vs. *negativo*. A partir de dicha clasificación, pueden calcularse grados de polaridad para temas formados por conjuntos de documentos, mediante el cómputo de elementos negativos y positivos en dichos temas.

Inicialmente, se había aplicado a analizar el contenido de textos extensos, como cartas, ensayos, emails, etc. Con la extensión del uso del Internet, nuevas plataformas y formatos de creación de contenido, ha ganado gran relevancia, sobre todo en el análisis de *blogging*, *microblogging* y en general, *posts* en redes sociales.

Como se ha comentado en la sección 2.1, el usuario actual hace uso de Internet para compartir su punto de vista, discutir sobre temáticas variadas, escribir opiniones sobre productos, recomendaciones, etc. y esto constituye una fuente extensa de información, tanto para la investigación como para la industria.

Encontramos pues, que el desarrollo de la técnicas han encontrado en *Twitter* un entorno perfecto en los últimos años y es el objeto de estudio de muchas de las aproximaciones en la literatura de los últimos años.

Sin embargo, existen otros muchos casos relevantes, como son los *reviews* de películas, de productos, noticias y blogs.

Podemos agrupar estas aproximaciones en dos categorías, ya se basen en:

- Análisis léxico

- Machine Learning

Las técnicas de análisis léxico se basan en el uso de un diccionario formado por *lexicons* pre-anotados. El texto que queremos analizar es *tokenizado* y los *tokens* obtenidos son comparados con el diccionario, uno por uno. Si existen coincidencias, aumentamos la cuenta de un *score* para el texto que se evalúa. Mientras que si no existen coincidencias, asignamos una anotación negativa a la palabra no encontrada, o decrementamos el *score* para dicho texto. Tras estas operaciones, se considera el *score* total como una medida del sentimiento del texto analizado.

Gran parte del esfuerzo de estas aproximaciones se ha empleado en estimar qué información léxica funciona mejor para esta tarea. Se ha conseguido una *accuracy* del 80 % en frases individuales, a través del uso de *lexicons* manualmente etiquetados, consistentes sólo de adjetivos, considerados claves para inclinar la posición del texto en una u otra dirección. Aplicando esta técnica a una base de datos de opiniones sobre películas, se obtuvo un *accuracy* de alrededor del 62 %. En esta dirección, (Kamps et al., 2004)(Hatzivassiloglou y Wiebe, 2000), se propuso reemplazar el lexicon por dos *queries* realizadas en un motor de búsqueda para cada palabra: (palabra+*good*) y (palabra+*bad*). El *score* se calculará con aquellas *queries* que devuelvan mayor número de resultados. Esta aproximación mejoró los resultados del experimento anterior del 62 % al 65 %. Posteriormente, comenzó a utilizarse la base de datos *WordNet* para calcular el *scoring*, comparando la palabra objetvio con las palabras pivote (*good* y *bad*) y calculando la mínima distancia de camino entre palabras en *WordNet* (Kennedy y Inkpen, 2006). Esta aproximación consiguió un 64 % de *accuracy* para el problema anterior.

En una dirección diferente, (Andreevskaia, Bergler, y Urseanu, 2007) evalúa el *gap* semántico entre las diferentes palabras sustrayendo el set de palabras positivos de las negativas, alcanzando un 82 % de *accuracy* para el mismo problema.

Este enfoque tiene varias limitaciones, entre las que podemos destacar la dependencia del contexto y la eficiencia. La primera es trivial y la segunda se deriva del tiempo y complejidad necesarias para realizar una evaluación, en la que el *accuracy* empeora en gran medida a medida que el número de palabras crece.

A pesar de que esta aproximación pueda parecer simplista o naïf, ha sido utilizada y comprobada como una aproximación válida y con resultados positivos durante mucho tiempo.

En otra categoría encontramos las técnicas basadas en *Machine Learning*, que han acaparado el interés de la investigación en la última década. De manera usual, se utilizan procedimientos de clasificación de aprendizaje supervisado, es decir, se construyen modelos que entrenan con datos previamente anotados en una u otra categoría. A continuación estos modelos infieren una categoría a datos sin categoría, que no han sido procesados previante en el entrenamiento.

La clave para construir dichos modelos de manera que sean eficaces para un caso particular es una correcta extracción, selección e ingeniería de *features* a partir del

texto. Algunos ejemplos de *features* que se utilizan son, por ejemplo, unigrams, bi-grams, tri-grams, longitud de muestra, etiquetado *Part-of-Speech*, diccionarios, frecuencia de palabras, etc. Otros factores también impactarán sobremanera en los resultados, como los pasos de preprocesado del texto (limpieza de carácteres extraños, mayúsculas, *stop-words*) Como modelos a destacar entontramos *Support Vector Machines* o *Naïve-Bayes*, con los que se consiguen *accuracies* de entre el 63 % y el 80 % dependiendo de las *features* utilizadas.

Las aproximaciones basadas en *Machine Learning*, (Java y others, 2007)(Durant y Smith, 2006)(Prasad, 2010), presentan diversos retos, como el diseño de un clasificador, la disponibilidad, sesgo y calidad de los datos o la ingeniería de *features* relevantes. Por otro lado, se superan los retos del enfoque léxico, en cuanto a capacidad de generalización, eficiencia y *accuracy*.

Como una subcategoría del *Machine Learning* que ha superado todas las expectativas, encotramos el *Deep Learning*, donde los modelos que se construyen están basados en redes neuronales. En el caso del Procesamiento del Lenguaje Natural, el introducir el concepto de representación vectorial del texto y los *word-embeddings* ha transformado completamente la manera de aproximar muchos problemas, entre ellos el de *Sentiment Analysis*. Esta innovación permite, capturar el contexto semántico del entorno de cada palabra del texto, además de eliminar el problema de la ingeniería manual de *features*.

Entre los modelos de *Deep Learning* utilizados en el problema de clasificación de texto y análisis de sentimiento encontramos relevante el uso de redes neuronales de tipo convolucionales (CNN) (Kim, 2014) y redes Long-Short-Term-Memory (LSTM) (Tai, Socher, y Manning, 2015) y sus derivaciones posteriores.

Las CNNs son un tipo de red neuronal usadada tradicionalmente en reconocimiento de imágenes, que también comenzó a usarse en el campo del Procesamiento del Lenguaje de Natural, a partir de la adopción del uso las representaciones vectoriales como *word2vec*. Este tipo de redes proveen muy buenos resultados, incluso en sus parametrización más básica, para datos secuenciales que tienen un patrón de jerarquía interna, descomponiendo en sus capas dicho patrón en patrones cada vez más y más pequeños. Las redes LSTM, por otra parte, son un tipo de red neuronal recurrente (RCN), también usada tradicionalmente en casos donde los datos tienen una estructura secuencial, como es el caso del reconocimiento de habla o de imágenes.

Se propone su uso para la clasificación de sentimiento en (Tai, Socher, y Manning, 2015), debido a su capacidad de preservar el valor semántico y a la hora de procesar la estructura estrutura sintáctica del texto. Tanto en (Kim, 2014) como en (Tai, Socher, y Manning, 2015) se consiguen valores de *accuracy* superiores al 85 %.

En nuestro caso, el análisis de sentimiento es un aspecto clave de nuestra propuesta de detección de controversia, ya que en nuestra definición, la polaridad de

las posiciones ante el tema de controversia es una condición necesaria, pero no suficiente de controversia. Este análisis de sentimiento, en definitiva de polaridad, en el ámbito de la detección de controversia está ampliamente respaldado por múltiples aproximaciones presentadas en 2.2, en ocasiones como técnica de base (Wang y Cardie, 2016) y en ocasiones como aspecto conmplementario (Garimella et al., 2018).

Por la relativa facilidad técnica con la que actualmente podemos configurar un modelo de *Deep-Learning* básico que nos garantice unos resultados aceptables, hemos considerado usar una representación *word2vec* de nuestro corpus para entrenar una red neuronal básica de LSTM como modelo de clasificación binaria para clasificar oraciones como *positivas* o *negativas*.

Capítulo 3

Método para la Detección de Controversia

En esta sección, y basándonos en las aproximaciones del estado del arte, exponemos nuestra propuesta consistente en una definición amplia de controversia así como un sistema para detectarla. Dicho sistema está compuesto por diferentes tareas, que serán presentadas de una en una, en orden de uso dentro del sistema. Como parte de esta exposición, se describirá el proceso de evolución del sistema hasta su estado actual a través de las diferentes decisiones que se han tomado y referenciando a los elementos extraídos del estado del arte.

3.1. Introducción

Como ya se ha mencionado, nuestra propuesta consta de dos partes:

1. Una definición de controversia, reuniendo diferentes elementos presentes en diferentes aproximaciones del estado del arte, que abarque un significado de amplio espectro y sea técnicamente aplicable.
2. Un sistema para detectar dicha controversia, basado en nuestra propuesta de definición, donde las diferentes tareas a ejecutar se corresponden con los requisitos que exigimos para poder considerar un escenario como controvertido.

3.1.1. Definición de Controversia

Hemos comprobado en la sección 2.2 que no existe un consenso sobre una definición amplia y explícita de controversia. La mayoría de las aproximaciones que hemos analizado utilizan conceptos heurísticos de controversia, implícitos al caso de uso concreto o a la plataforma de la que obtienen los datos, como pueden ser *Twitter*,

Wikipedia, etc. En otros casos, comprobamos que la definición se basa enteramente en un único aspecto de la controversia, como puede ser la polaridad, el desacuerdo o la topología de grafo.

El primero de nuestros objetivos a cumplir será, por tanto, definir y concretar qué significa controversia, qué implica ser controvertido y en qué consiste exactamente la tarea de su detección.

Como primera aproximación a la definición de controversia, aglutinamos las ideas generales presentes en gran parte del trabajo previo, y revisando la definición proporcionada por los diferentes diccionarios de la lengua, podemos sintetizar que:

> "controversia es aquella situación de discusión pública (y reiterada) sobre un tema que genera posiciones contrapuestas en los grupos de individuos que participan en dicha discusión."

Es importante que esta definición no solo tenga significado técnico, sino que también preserve el significado original de lo que entendemos por controversia en el ámbito general. Pero necesitamos también que sea de carácter amplio, aplicable para diferentes contextos y también implementable técnicamente. En ese sentido, tendremos que matizar qué significan conceptos como *discusión*, *grupos*, *contraposición*, etc. en nuestra propuesta.

Construiremos pues una definición basándonos en unas condiciones mínimas compatibles para poder considerar las diversas situaciones, descritas con anterioridad, susceptibles de ser analizadas en busca de controversia.

Basamos nuestra definición en los siguientes conceptos y asunciones:

1. El **objeto de controversia** será un tema concreto, claramente descrito, y un conjunto de elementos textuales sobre ese tema, generados por individuos o grupos de individuos. La matización en la clara descripción y concreción del tema tiene como objetivo minimizar la ambigüedad, que existirá siempre en cierta medida, con la que diferentes individuos entienden dicho tema.

   Dicho tema no se proporciona de manera explícita al sistema, sino que está implícitamente definido y representado por el conjunto de documentos de texto proporcionados, que es la única información usada por el sistema para su evaluación de controversia.

   Concretamente en nuestro caso, el objeto de controversia es un caso de uso para un fármaco, es decir, la combinación de fármaco y condición médica, y un conjunto de comentarios de usuarios generados a raíz de ese caso de uso.

2. La **granularidad de la controversia**: la controversia puede darse a diferentes niveles, dependiendo del tipo de documentos que se consideren y de su estructura.

   Si queremos determinar la controversia de un documento concreto, deberíamos aplicar nuestra definición a nivel de elementos internos del texto. Es decir,

un único usuario podría aportar argumentos a favor y en contra del tema del documento, y esos serían nuestros elementos a considerar. Sin embargo, si queremos conocer la controversia de un tema en *Twitter*, los elementos lógicos a considerar podrían ser los diferentes *tweets* que provienen de diferentes usuarios que publican con un mismo *hashtag.*

3. La **presencia de argumentación** en los elementos textuales (en nuestro caso, los comentarios) es una condición necesaria pero no suficiente para que se dé el escenario de controversia. El aspecto de *discusión* presente en la controversia consiste en la exposición de diferentes argumentos, cuyo objetivo o contenido puede ser analizado posteriormente.

   Esta condición, a modo de filtro, nos permite distinguir entre aquellos elementos que han sido elaborados sobre una base racional de aquellos que constituyen solamente una expresión de un sentimiento. Como ejemplo, pretendemos distinguir un *"¡Messi es el mejor!"* de un *"Después de haber fallado solamente 10 ocasiones de 15 posibles, Messi es claramente superior a los demás"*. Debemos destacar que en el marco de este trabajo, estamos interesados solamente en la presencia o indicios de argumentación, ya que nuestra intención es la de proponer una definición de condiciones mínimas. Es parte del trabajo futuro, aplicar un análisis más profundo de esta argumentación, como extraer sus componentes o analizar si son razonamientos lógicamente correctos.

4. La **composición interna de la controversia**: al igual que ocurre con la detección de argumentación, existirá una estructura interna en el fenómeno de la controversia: qué elementos interactúan con qué otros elementos, qué argumentos apoyan a otros argumentos y están en contraposición a otros argumentos, etc.

   En nuestra propuesta, como una primera aproximación amplia a este problema, no tendremos en cuenta ni analizaremos las características de esta estructura. De igual manera, para nosotros los argumentos serán la unidad básica que extraer del texto, sin importar a qué usuario pertenecen, ni la relación entre usuarios ni entre argumentos, más allá de la posibilidad de agruparlos por similitud. Explorar esta composición interna es un potencial paso futuro en la línea de ampliar esta propuesta.

5. La **agrupabilidad de los argumentos**: para que exista controversia, los diferentes argumentos deben ser susceptibles de ser agrupados. Estos grupos se considerarán *sub-temas* o *aspectos* de la controversia, sobre los que los individuos están posicionándose. Es decir, para que exista enfrentamiento entre posiciones, las posiciones deben poder formarse en base a grupos temáticamente similares.

6. La **polaridad de los grupos**: los grupos deben estar caracterizados por una polaridad, que puede ser medible mediante técnicas como el *sentiment*

analysis. A través de esta medida de polaridad, identificaremos cuál es la posición neta de los diferentes grupos, es decir, de las *sub-temáticas* hacia el tema.

7. La **confrontación de los grupos**: para que se dé controversia, debe existir confrontación entre grupos. Entedemos confrontación como la existencia de grupos de tamaños relevantes y comparables, que poseen polaridades opuestas.

   Un escenario hipotético de gran controversia sería la formación de dos grupos de igual tamaño con polaridades internas netas totalmente inclinadas hacia posiciones opuestas. Un caso de baja controversia consistiría en un escenario con un gran grupo que aglomera gran parte de los elementos con una polaridad definida, y por otro lado diversos grupos de pequeño tamaño de polaridad opuesta.

8. La **cuantificación de la controversia**: dadas las propiedades, condiciones y asunciones anteriores, se espera que la controversia, tal como la hemos definido pueda cuantificarse, y que esta cantidad tenga un valor y significado coherente para un evaluador humano que comparta dicha definición.

9. La **dependencia del contexto**: o de manera trivial, dependiente del conjunto de elementos textuales que se consideren para su estimación. Es una característica implícita de la controversia a tener presente, sobre todo en la interpretación y evaluación por un evaluador humano.

   Los elementos textuales serán procesados de manera indistinguible, pero hay que tener en cuenta que si quisiéramos estimar la controversia de un tema, seleccionando para ello un conjunto de documentos, dicho conjunto representará un contexto temporal, sociológico, geográfico, etc. Como ya se ha mencionado, si queremos conocer cuál es el grado de controversia *del aborto*, habría qué preguntarse dónde y cuándo para hacer una selección adecuada de documentos.

Basándonos en estas condiciones y asunciones, podemos derivar una definición un poco más concreta que la expuesta inicialmente:

Definición 1. *Dado un tema de discusión determinado, y un conjunto de elementos de contenido textual generados por una comunidad de individuos en relación a dicho tema, se considerará que tal tema posee un grado medible de controversia respecto a tal comunidad, si:*

1. *Puede extraerse un conjunto suficiente de unidades textuales con indicios de argumentación, a los que llamaremos argumentos.*

2. *Dichos argumentos pueden agruparse en un conjunto de grupos, que constituirán las temáticas sobre las que se apoyan las posiciones.*

3. *Dichos grupos poseen una polaridad que puede ser evaluada.*

4. *Grupos de tamaño relevante y comparable entre sí manifiestan polaridades enfrentadas.*

Esta definición nos aporta una base sobre la que poder construir nuestro proceso de detección de controversia, pero también nos queda claro que *a priori*, no obtenemos una respuesta binaria a la pregunta de si un tema es controvertido o no, ya que los diferentes puntos sobre los que construimos nuestra definición generan la necesidad de cuantificar la controversia con un valor que nos aporte información sobre cuán controvertido es un tema.

3.1.2. Propuesta de Sistema para la Detección de Controversia

Partiendo de la definición amplia proporcionada en el apartado anterior, podemos distinguir claramente cuatro tareas diferentes que abordar, una por cada punto de la definición, como se muestra en la figura 3.1 y que se concretan en:

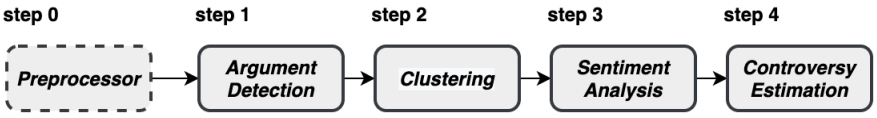

Figura 3.1: Representación esquemática de los cuatro pasos principales de la propuesta, basados en los puntos de la definición.

1. **Detección de Argumentos**: considerado como un problema de clasificación binaria para nuestras unidades de texto en "argumentativo" vs. "no-argumentativo".

2. ***Clustering* de Argumentos**: aplicación de técnicas de *clustering* para obtener un conjunto de grupos de argumentos similares semánticamente.

3. **Análisis de Sentimiento**: estimación binaria genérica de sentimiento para cada unidad de texto ("positivo" vs. "negativo").

4. **Estimación de Controversia**: utilizando la polaridad y el tamaño de los *clusters*, derivación y aplicación de una función de estimación cuantitiva de controversia para cada tema.

Para cada uno de los pasos mencionados, será necesario diseñar un componente que cubra los requisitos que esperamos para nuestro caso. Nuestra aproximación se basa en la consideración del problema de detección de controversia como un tipo de problema de detección de argumentos, posteriormente enriquecido con la extracción

de *clusters* y análisis de polaridad para finalmente realizar una estimación cuantitiva de controversia.

Como tal, debemos destacar el paso de la detección de argumentos como el más importante, ya que sirve de base para el resto del proceso. Por tanto, durante el desarrollo del caso de estudio en el que aplicaremos nuestra propuesta de sistema detección, centramos nuestros esfuerzos y recursos en asegurar unos resultados de clasificación sólidos para esta tarea. Para los pasos de clustering y análisis de sentimiento, aplicamos una metodología genérica que nos garantice una funcionalidad suficiente para nuestros objetivos, ya que el estado estado del arte en estas dos tareas está muy consolidado y sus resultados son satisfactorios y extrapolables a cualquier dominio.

La evaluación final se realizará sobre los resultados obtenidos del último paso, el estimador de controversia, que incorpora las salidas de los pasos anteriores. Un grupo de evaluadores humanos, a los que se les proporcionará la definición *no-técnica* de controversia realizarán una evaluación del dataset de evaluación seleccionado, para compararlos cualitativamente con los datos proporcionados a la salida del sistema.

3.1.3. Caso particular: comentarios de usuarios sobre fármacos en foros del ámbito médico.

Junto a nuestra propuesta amplia para la controversia y su detección, evaluamos su aplicación práctica a través del caso concreto de comentarios de usuarios sobre fármacos en foros de esta temática. Para ello utilizaremos el corpus ***Drug Review Dataset***, presentado en (Gräßer et al., 2018).

Dicho corpus consiste en 215,063 comentarios de usuarios sobre fármacos específicos, así como la afección para la que son usados y un *rating* del 1 al 10 sobre la satisfacción del usuario, además de diversos metadatos. Para la aplicación de nuestro sistema, utilizaremos sólamente los datos proporcionados por los comentarios, nombre del fármaco y afección. Adicionalmente, se utilizará en pasos posteriores (sección 3.2.4) el valor del *rating* como un etiquetado de la polaridad del comentario, ante la falta de recursos para realizar un etiquetado propio para esa tarea.

Consideraremos un *caso de uso*, formado por la combinación de fármaco y afección, como el tema y objeto potencial de controversia. Por ejemplo: *paracetamol+cefaleas*. Cada caso de uso estará constituido por un conjunto de comentarios, siendo cada comentario un documento para el sistema. A continuación, son segmentados y procesados a nivel de oraciones individuales.

Cada oración será: *(1)* clasificada como "argumentativa" o "no-argumentativa"; *(2)* agregada a un *cluster* con otras oraciones semánticamente similares, que constituirán subtemas o aspectos de la controversia (como puede ser el caso de un efecto secundario concreto);*(3)* clasificada de acuerdo a su sentimiento ("positi-

vo","negativo"), y *(4)* finalmente, se aplica la estimación de controversia sobre el conjunto de *subtemas* para el caso de uso.

Como resultado tendremos una cuantificación relativa para ese caso de uso de fármaco, comparable con otros casos de uso.

3.2. Descripción del Sistema

A continuación, realizaremos paso por paso una descripción de los componentes concretos del sistema, así como el proceso de su diseño y desarrollo. En la figura 3.2 encontraremos la arquitectura detallada del sistema implementado, basada en la estrucura presentada de manera esquemática en la sección 3.1.

3.2.1. Preprocesado

El primer paso aplicado a todo texto que vaya a ser procesado en el sistema es una estandarización y limpieza previa. En nuestro caso, este preprocesado consiste en: tokenización, expansión de contracciones (*I don't know* pasa a ser *I do not know.*), conversión mayúsculas a minúsculas, lematización (preservando también una versión del texto sin lematizar) y sustitución de dígitos. Se han preservado *stop-words* y signos de puntuación, ya que en nuestro caso juegan un papel relevante, por su presencia en diferentes *features* para la detección de argumentación.

A continuación, se realiza un segmentado por oración utilizando la funcionalidad de la librería *SpaCy*[1], que tiene en cuenta casos más complejos que considerar únicamente simples signos de puntuación.

Como salida, tendremos dos versiones del texto preprocesado, una versión lematizada y otra sin lematizar, ya que ambas nos serán de utilidad en diferentes pasos de la extracción de *features*.

3.2.2. Detección de Argumentos

El primer componente funcional del *pipeline* es el detector de argumentos. En nuestra propuesta, éste consta de dos partes principales: la clasificación genérica de argumentos, utilizando un enfoque generalista y un segundo clasificador con características de dominio.

Será este, por tanto, un punto donde el proceso puede ser adaptado en profundidad para un caso concreto, si se tienen una serie de *features* para tal dominio, complementando la señal original del clasificador genérico.

En nuestro caso, combinamos ambas señales y una señal extra de dominio, que consideramos empíricamente relevante, mediante un voto mayoritario del que surge

[1] https://spacy.io :librería *open-source* que incluye diferentes funcionalidades y herramientas para tareas de Procesamiento del Lenguaje Natural.

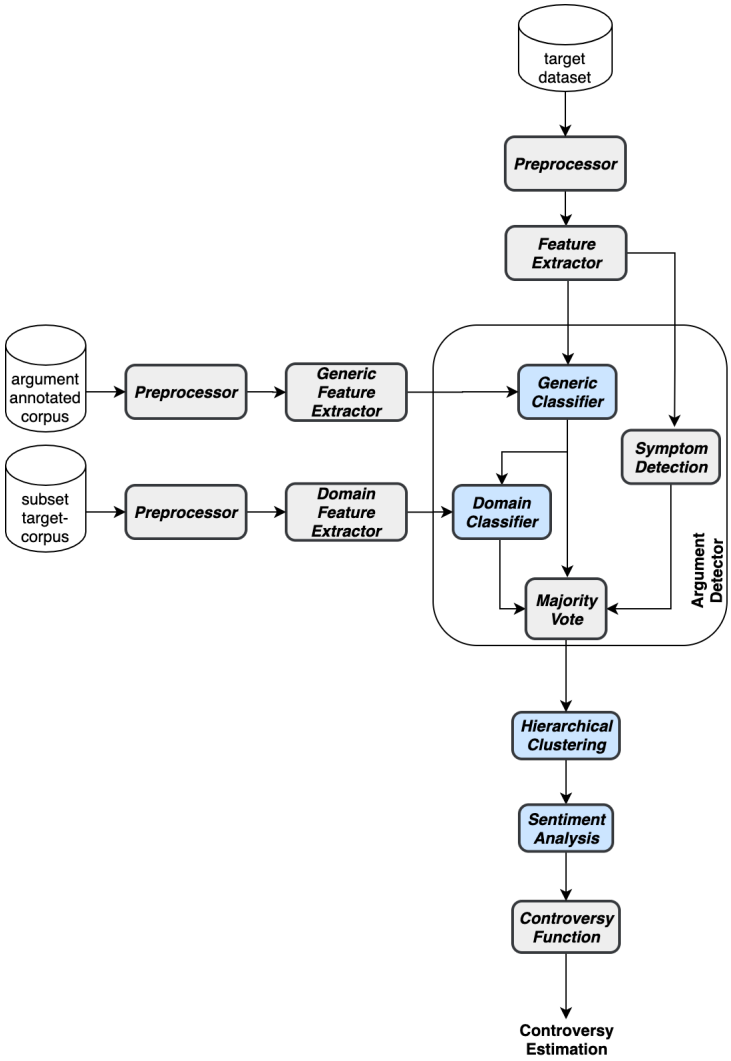

Figura 3.2: Estructura del *pipeline* usado en nuestro sistema. Verticalmente encontramos la ejecución del *pipeline* mientras que horizontalmente encontramos el proceso de entrenamiento de los modelos de clasificación para detección de argumentos. Puede identificarse en color azul aquellos componentes que contienen modelos estadísticos, y en color gris, elementos funcionales definidos por nuestras decisiones técnicas y definiciones.

la señal final de clasificación binaria.

Analizamos y evaluamos, paso por paso, sus componentes.

3.2.2.1. Clasificador Genérico

Para la construcción de un clasificador genérico de argumentación, seguiremos las pautas que hemos extraído a partir del estado del arte. Seguimos una metodología similar a la utilizada en (Stab y Gurevych, 2017) y (Stab y Gurevych, 2014b), a través de la cual, mediante una serie de *features* consideradas relevantes para detectar mecanismos propios de la argumentación, y utilizando un corpus previamente anotado, se entrena un modelo de aprendizaje máquina para construir un clasificador de unidades de texto.

Sin embargo, tendremos que adaptar la metodología presentada en (Stab y Gurevych, 2014b) por dos razones principales. En primer lugar, la clasificación se utiliza para identificar sub-componentes de la argumentación como *evidences*, *claims*, *major-claims*, y por tanto la anotación se ha realizado sobre unidades inferiores a la oración. En segundo lugar, se usa dicha metodología sobre el corpus de *Persuasive Essays* (Stab y Gurevych, 2014a), consistente en una serie de textos de tipo ensayístico, con características estructurales y contextuales bien definidas.

Nuestra hipótesis para esta tarea es que las *features* propuestas para resolver la tarea son también válidas para identificar la argumentación a nivel de oración, simplificando el problema para dos categorías: *argumento* vs. *no argumento*. Para probar dicha hipótesis, debemos aplicar las *features* propuestas por (Stab y Gurevych, 2017) para construir un clasificador de argumentación binario ("argumento" vs. "no-argumento") sobre un corpus propiamente anotado en ese sentido, a nivel de oración y para diferentes dominios. Si obtenemos resultados satisfactorios para dicha clasificación, podemos considerar que el uso de las *features* propuestas es válido para nuestro caso.

Selección de features

Como primer paso, debemos seleccionar qué *features* de las presentadas en (Stab y Gurevych, 2017) (tabla 3.1) pueden sernos de utilidad y cuáles deben descartarse, por la dependencia con el caso de uso utilizado en dicho trabajo.

Podemos observar en la tabla 3.1 que las *features* están diseñadas para extraer información a diferentes niveles (término, oración, párrafo, documento, etc.) y en diferentes aspectos (léxico, estructural, contextual, sintáctico, semántico, etc.)

Encontraremos en dicha tabla, que las descripciones se refieren tanto a *features* obtenidas a nivel de oración como a nivel de *"componente"*. Esto se debe a que originalmente en (Stab y Gurevych, 2017), la unidad considerada es la de "componentes" de la argumentación, que son a menudo inferiores a la oración. En nuestra adaptación, debemos considerar "componente" y "oración" como equivalentes.

En nuestro caso, tenemos un sistema que opera a nivel de oración, conservando información sobre el comentario original al que dicha oración pertenece y pudiendo extraer de él ciertas *features* (por ejemplo, estadísticas de *tokens* de la oración

anterior y posterior). No obstante, dicho comentario, en línea con el diseño independiente del caso de nuestro sistema, no puede equipararse a un párrafo de un texto, por no poder garantizar una estructura regular ni una coherencia semántica equiparable a un párrafo en un texto.

En ese sentido, nos vemos en la posición de descartar aquellas *features* (marcadas en rojo en la tabla 3.1) que hacen referencia a relaciones entre párrafos, como son las *features* de tipo *"Contextual"* o la *feature "Posición del componente"*. Debemos descartar también aquellas *features* basadas en la estructura interna del argumento, como es la *feature* de tipo *"Probabilidad"*, basada en la clasificación de componentes de argumentación (*claim*, *major-claim* y *evidence*). Igualmente, la *feature* de tipo *"Discursivo"* se basa en el grado de superposición de diferentes relaciones entre los componentes, sirviéndose de un *parser* basado en PDTB[2].

Debe tenerse en cuenta que nuestro objetivo es el de detectar de manera binaria si una oración es de carácter argumentativo o no, por lo que se reduce en gran medida la complejidad y granularidad de las *features* a utilizar. Analizando una serie de instancias potencialmente argumentativas, correspondientes a diferentes corpus mencionados en el estado del arte sección 2.3.1, se extrae la observación de que, para construir un detector de argumentación de carácter genérico, los aspectos sintácticos, estructurales y una serie de indicadores pueden constituir las *features* más relevantes, seguidas a continuación de los *word-embeddings*, y en menor medida las *features* léxicas, debido a la dependencia del dominio.

Encontramos la mayoría de estos aspectos bien representados en el resto de *features* de la tabla 3.1. A nivel sintáctico, parece evidente que una mayor complejidad de oración (número de sub-oraciones, profundidad del árbol sintáctico, modo del verbo, etc.) representará una idea más compleja y potencialmente más argumentativa (elementos causa-efecto, pro-contra, dependencias entre conceptos, etc.) que oraciones más simples.

Los indicadores, tanto de primera persona usados para expresar puntos de vista, como los presentados en la lista del apéndice C1, están presentes como nexos entre ideas y reguladores del flujo del discurso. Por ejemplo: *"consequently"*, *"moreover"*, *"even though"*, etc. Estos indicadores son una *feature* simple, pero efectiva y generalizable, común a todos los dominios en los que se pueda dar argumentación.

Las estadísticas de *tokens* nos resultan igualmente de interés, de nuevo como una posible medida de complejidad de la oración. Respecto al uso de *embeddings*, consideramos que puede resultar de interés, aún en dependencia del dominio, incluir información contextual en nuestras *features*. Para dicha *feature*, hemos usado los *word-embeddings* presentados en (Baroni, Dinu, y Kruszewski, 2014).

Tras incluir todas las *features* mencionadas, tenemos la impresión de que podemos capturar aún más características interesantes sobre los diferentes patrones de estructuras sintácticas que pueden darse en las instancias argumentativas. Propone-

[2] https://www.seas.upenn.edu/~pdtb/ Penn Discurse Treebank

mos una *feature* adicional, para intentar abarcar esos posibles patrones, presentada en la tabla 3.2. Mediante esta *feature*, realizamos una representación en POS-*tags* de los términos presentes en la oración y aplicamos *tf-idf*[3] para los *unigrams*, *bi-grams* y *tri-grams* más frecuentes. De esta manera tendremos en cuenta en nuestro modelo, ocurrencias de combinaciones de POS-*tags* relevantes en los documentos, junto a las ya incluídas estructura del árbol sintáctico, su profundidad, y la distribución de POS-*tags* en las *features* mencionadas.

Corpus de entrenamiento

Una vez seleccionadas las *features*, necesitamos un corpus adecuado a nuestras necesidades. El corpus *Persuasive Essays* (Stab y Gurevych, 2014a) utilizado en la aproximación que propone nuestras features de referencia (Stab y Gurevych, 2014b), como hemos mencionado, está anotado a nivel de unidades de texto inferiores a la oración con las etiquetas propias de los sub-componentes de la argumentación.

El corpus consiste en documentos de tipo ensayístico que defienden o atacan una idea principal, con una estructura muy parecida y definida ("Introducción", "Presentación de las tesis", "Conclusión"). Estructuralmente, este corpus no es un caso muy generalizable, ya que introduce implícitamente la estructura interna de un documento típico, como hemos descrito y que nos han hecho descartar algunas de las features de la tabla 3.1 en la sección anterior. Otra desventaja es el hecho de que esté anotado a nivel de sub-componente de argumentación (*major-claim*, *claim*, *evidence*), mientras que para nuestro caso, estamos interesados en la detección a nivel de argumento.

Como hemos visto en la sección 2.3.1, el corpus que más se adapta a nuestras necesidades es *Cross-Domain Sentential Argument Mining* descrito en el artículo (Stab et al., 2018), consistente en 25,000 oraciones etiquetadas como *argumentos-pro* , *argumentos-contra* y *no-argumentos*, distribuidos en ocho temas de diferentes ámbitios considerados como controvertidos. Debido a problemas técnicos en el uso del software proporcionado por (Stab et al., 2018) para generar el dataset, se han generado solamente seis temas de los ocho posibles, con un total de 19,443 oraciones. [4]

Este dataset constituye para nosotros la mejor opción debido a su carácter cross-dominio, que nos permitirá no solo construir un clasificador que no dependa en gran medida del dominio, sino también trabajar a nivel de oración y por la granularidad de las anotaciones, ya que en la mayoría de los *corpora* encontrados en el estado del arte, en los que las anotaciones se realizaban a un nivel de sub-componente de la argumentación (*evidence*, *claim*, etc.) en lugar de a nivel de argumento (*argument*

[3] Term Frequency - Inverse Document Frequency

[4] Por cuestiones de recursos y desarrollo del trabajo, no se ha investigado en profundidad este fallo técnico.

Tipo	***Feature***	***Descripción***
Léxico	Unigrams	Unigrams binarios y lematizados del componente y sus tokens precedentes
	Tuplas de dependencia	Tuplas lematizadas del árbol de dependencias (2k más frecuentes).
Estructural	Estadísticas de tokens	Número de tokens del componente, número de tokens de los componentes anterior y posterior.
	Posición del componente	El componente está al principio o al final del párrafo; si está presenta en la introducción o conclusión; posición relativa en el párrafo; número de componentes precedentes y siguientes en el párrafo.
Indicadores	Indicadores de tipo	Indicadores de avance, retroceso, tesis o refutación presentes en el componente o en sus *tokens* precedentes. (lista completa en el apéndice, tabla C1).
	Indicatores de primera persona	"I", "me", "my", "mine", o "myself" presentes en el componente o en los *tokens* precedentes.
Contextual	Indicadores de tipo en el contexto	Indicadores de avance, retroceso, tesis o refutación a continuación o previos al componente en su párrafo.
	Frases compartidas	Predicados nominales y verbales compartidos por la introducción o conclusión.
Sintáctico	Sub-oraciones	# sub-oraciones en la oración.
	Profundidad del árbol sintáctico	Profundidad del árbol sintático de la oración.
	Tiempo verbal	Tiempo verbal del verbo principal del componente.
	Verbos modales	Verbos modales presentes en el componente.
	Distribución POS	Distribución POS *tags* en el componente
Probabilidad	Probabilidad de tipo	Probabilidad condicional de que el componente sea un *claim*, *major claim* o *premise*, dados sus *tokens* precendentes.
Discursivo	Tripletes de discurso	Relaciones de discurso PDTB coincidentes con el componente.
Embeddings	*word-embeddings*	Suma de los vectores de todas las palabras que forma el componente y sus *tokens* precedentes.

Tabla 3.1: *Features* propuestas para detección de argumentos, propuestas en el estado del arte. En rojo, *features* no utilizadas por criterio propio.

vs. *no-argument*).

Consideramos la clasificación como binaria, para lo cual, aglutinamos las clases *argumento-pro* y *argumento-contra* bajo una misma clase *argumento*.

Tipo	***Feature***	***Descripción***
Patrones estructurales	Frecuencia de POS *tags*	Frecuencia de unigrams, bi-grams y trigrams de POS, sustituyendo cada *token* por su POS *tag* y extrayendo su tf-idf para los unigrams, bigrams y trigrams más frecuentes.

Tabla 3.2: *Feature* de propuesta propia para detección de argumentos.

Modelos de aprendizaje máquina

Se han entrenado diversos modelos lineales de aprendizaje máquina basados en un optimizador *Stochastic Gradient Descent* con diferentes funciones de pérdida: *Support Vector Machine (SVM)*, *Huber*, *Modified Huber* y *Regresión Logística*) como algunos de los usados en el estado del arte, para los que se ha ejecturado un *grid-search* de parámetros óptimos.

Evaluación

Nuestro objetivo es el de validar el uso de las *features* descritas por la tabla 3.1 sobre el corpus *Cross-Domain Sentential Argument Mining* (Stab et al., 2018). Originalmente, dichas *features* han sido utilizadas para detección de sub-elementos de la argumentación en el corpus *Persuasive Essays* (Stab y Gurevych, 2017). Pero, como se ha mencionado, el corpus *Cross-Domain Sentential Argument Mining* (Stab et al., 2018) se adapta mejor a nuestro desarrollo y necesitamos, por tanto, conocer si dichas *features* siguen siendo válidas para este caso, a pesar de que este último corpus sea cross-dominio y anotado a nivel de oración.

Consideramos las 19,443 instancias del corpus *Cross-Domain Sentential Argument Mining* (Stab et al., 2018), en una división de sets para entrenamiento (70 %) y test (30 %), con un *cross-validation* de 5x2 y una clase mayoritaria del 56 % sobre el total de instancias, que usaremos como *baseline* no-informativo.

Utilizando los modelos de aprendizaje máquina descritos en la sección anterior, obtenemos el mejor resultado para una configuración de *Regresión Logística* con optimizador *Stochastic Gradient Descent (SGD)*, con un $F1$ =.73, superior a nuestro *baseline* no-informativo de .56 y a los resultados presentados en el estado del arte (Stab et al., 2018) de $F1$ =.67.[5]

Con estos resultados, podemos considerar dichas *features* válidas para construir un clasificador binario de argumentación basado en el corpus considerado *Cross-Domain Sentential Argument Mining* (Stab et al., 2018).

Consideramos la configuración descrita como nuestra configuración final para este Clasificador Genérico.

[5] más detalle en la sección de resultados para este componente 4.1.1

El siguiente paso será evaluar los resultados de clasificación sobre el dataset *Drug Review Dataset* (Gräßer et al., 2018) de nuestro caso particular, descrito en la sección 3.1.3.

Para ello, extraemos un subconjunto de 5526 instancias del dataset de fármacos y aplicamos el clasificador genérico de argumentación.

Al no estar anotado, debemos realizar una evaluación manual, en primer lugar cualitativa, sobre los resultados de aplicar el clasificador entrenado con datos controvertidos multidominio a este caso.

Las conclusiones de nuestro análisis cualitativo son tales que, a pesar de ser existir una base de acierto, existe un gran número de casos que se consideran argumentativos en este dominio y que no están siendo correctamente clasificados, en lo que parece ser un escenario de alta precisión pero muy bajo *recall*.

Se observa que existe un *gap* conceptual entre lo que pueda significar *argumentación* en el corpus de entrenamiento (*Cross-Domain Sentential Argument Mining*) y nuestro caso particular (*Drug Review Dataset*), en cuanto a *features* que caracterizan la argumentación en el corpus de fármacos. Este *gap* se traduce en la exclusión de muchos casos considerados como *argumentativos* en este dominio.

Se nos presenta la opción, dados los recursos para realizar esta tarea, de incluir en nuestra arquitectura un componente de adaptación al dominio, que complemente la falta de generalización del modelo genérico a un caso tan particular como el nuestro.

3.2.2.2. Clasificador de Dominio

Para el desarrollo del clasificador de dominio, podemos distinguir tres fases consistentes en la realización de tres experimentos, esquematizados en la figura 3.4, utilizando los datasets descritos en la tabla 4.2, que han sido extraídos a partir del dataset *Drug Review Dataset*, como se muestra en el esquema de la figura 3.3.

Selección de features

En primer lugar, como paso previo, se opta por realizar un análisis cualitativo de los resultados de clasificar 1000 instancias a través del clasficador genérico en su configuración final, descrito en la sección anterior, para construir el dataset *train_1_dev*. A continuación, se ha realizado la anotación del dataset *train_1_ann*, teniendo en cuenta aquellas características que estuvieron presentes en los casos considerados como argumentativos por los anotadores. A partir del análisis cualtitativo de la argumentación en los dos pasos descritos, observamos que existen una serie de *features* de dominio que pueden usarse potencialmente para capturar todos aquellos casos de argumentación no clasificados por el clasificador genérico.

Dichas *features*, descritas en la tabla 3.4, consisten, en primer lugar, en entidades del ámbito médico, ya consideradas como relevantes en trabajos como (Carrillo-de

Exp.	Dataset	Instancias	Descripción
-	*pharma_dev*	68718	extraído aleatoriamente de *Drug Review Dataset*.
-	*dev*	5526	extraído aleatoriamente a partir de *pharma_dev*.
1	*train_1_dev*	1000	extraído aleatoriamente a partir de *dev*, clasificado por el Clasificador Genérico en su configuración final.
1, 2	*train_1_ann*	296	extraído aleatoriamente a partir de *dev* y anotado manualmente. Usado para crear *train_1* y para entrenar *classifier_dom_2*
1	*train_1*	1296	combinación de *train_1_dev* y *train_1_ann*, usado para entrenar *classifier_dom_2*.
2	*eval_1_ann*	419	extraído aleatoriamente de *dev* y anotado manualmente. Usado para evaluar *classifier_dom_1* y *classifier_dom_2*.
3	*train_2*	715	combinación de *train_1_ann* y *eval_2_ann*. Usado para entrenar *classifier_dom_3*.
3	*eval_2_ann*	215	extraído aleatoriamente de *dev* y anotado manualmente. Usado para evaluar *classifier_dom_3*.

Tabla 3.3: Datasets utilizados para la construcción del Clasificado de Dominio. El proceso de extracción de dichos datasets puede verse en la figura 3.3.

Albornoz et al., 2019). Dichas entidades pueden ser fácilmente extraídas a través de la herramienta *MetaMap*[6]. Entre ellas podemos encontrar ocurrencias de síntomas/efectos secundarios, sustancias, actividades relacionadas con el ámbito médico y verbos funcionales de este ámbito.

Tras nuestro análisis, comprobamos que existen mecanismos de argumentación del usuario, a la hora de adoptar una posición respecto al fármaco, en los que se describen experiencias, bien como un cambio de estado tras tomar el fármaco, bien como experiencias previas, bien como el proceso que siguió durante el uso del fármaco, etc.

En la misma dirección, construímos tres listas propias de indicadores, que resultan relevantes para la argumentación, referidos a diferentes aspectos de dicha argumentación: indicadores evolutivos, temporales y personales, no cubiertos completamente por las *features* anteriores. Los indicadores de carácter evolutivo describen ideas como procesos, causalidad, resultados, consecuencias, etc. que parecen usarse como razón de posicionamiento del usuario a favor o en contra del uso del fármaco. De manera análoga a las entidades extraídas para el ámbito médico, hemos considerado que mencionar elementos temporales, como unidades de tiempo,

[6] https://metamap.nlm.nih.gov/:*MetaMap* es una herramienta de uso público para reconocimiento de entidades del ámbito médico en textos mediante el uso del Metathesaurus de *UMLS (Unified Medical Language System)*

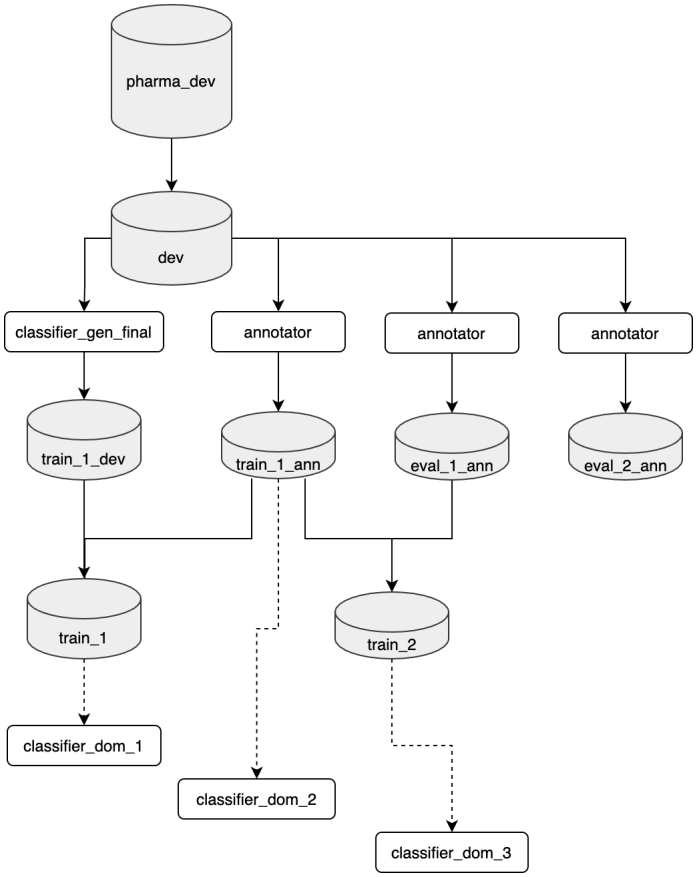

Figura 3.3: Descripción del proceso de construcción de datasets (tabla 4.2) para los experimentos descritos en la figura 3.4. Las líneas continuas representan transformación o extracción de datos, mientras que las discontinuas representan entrenamiento de un modelo. El proceso se ha desarrollado de izquierda a derecha y de arriba a abajo.

regularidad, adverbios temporales, etc. indican, a menudo, el hilo de una argumentación. Siguiendo con la analogía, mencionar otros sujetos, como familiares, amigos o especialistas médicos ayudan a construir la descripción de una experiencia y de una argumentación.

Estos indicadores léxicos nos son de gran utilidad, sobre todo para capturar experiencias de usuarios, que en este dominio se entremezclan con la argumentación. Sin embargo, no cada experiencia constituye un argumento, ya que aquellos comentarios que tienen un carácter exclusivamente expositivo, no pueden considerarse como argumentativos, al no poder garantizar que la exposición realizada tiene en realidad una intención de apoyo o descrédito del uso del fármaco. Por ejemplo, una oración del tipo *"My father uses this medicine for 2 years now..."*, tiene elementos

Tipo	***Feature***	***Descripción***
Ámbito Médico (MetaMap)	# síntomas	número de síntomas en la oración.
	# sustancias	número de sustancias en la oración.
	# cantidades	número de especificaciones numéricas.
	# actividades	actividades relacionadas con el ámbito médico (*improvement*, *intake*, *treatment*, etc.).
	# verbos funcionales	verbos relacionados con el ámbito médico (*take*, *digest*, *apply*, etc.).
Indicadores	Evolutivos	términos que indican una transición, cambio o evolución entre estados de salud (*change*, *better*, *worse*, *deteriorate*, *effect*, etc.).
	Temporales	términos que indican un aspecto temporal o de asiduidad en la oración (*day*, *hour*, *once*, *soon*, *normally*, etc.)
	Personales	términos que presentan a otros sujetos personales en la oración (*doctor*, *husband*, *partner*, *physician*, etc.).
Argumentativo	Confianza del clasificador genérico	La confianza en el intervalo [0,1] de que el elemento pertenezca a la clase "argumentativo", provisto por el primer clasificador.
Genérico	Longitud oración	número de caracteres de la oración.

Tabla 3.4: *Features* consideradas como relevantes para la detección de argumentos en el dataset *Drug Review Dataset*, tras el análisis cualitativo realizado en la sección anterior (lista completa de indicadores en el apéndice, tabla C2).

expositivos e indicadores como los comentados, pero no se trata de un argumento.

Esta es una de las razones por las que consideraremos que la mejor solución puede ser la combinación de las señales del clasificador genérico y las de un clasificador de dominio basado en este tipo de *features*, que cubran ambos aspectos de nuestra tarea. Esto se discutirá en la siguiente sección.

Para concluir, se han considerado además dos *features* que no pertenecen al dominio, pero aportan información útil de carácter general para construir nuestro modelo. Por un lado, incluimos la confianza de pertenecer a la clase *Argument*, provista por el clasificador genérico, aportando una medida del grado de argumentación genérica a nuestro set de *features*. Y por otro lado, incluimos la longitud de la oración, ya que es mucho menos probable detectar argumentación en oraciones cortar frente a oraciones largas.

El siguiente paso será construir un modelo que nos aporte resultados aceptables para esta componente de dominio. Para ello realizamos los siguientes experimentos[7],

[7]Los resultados en detalle pueden consultarse en la sección de Resultados para este componente, sección 4.1.2

esquematizados en la figura 3.4, sobre los datasets descritos en la figura 3.3.

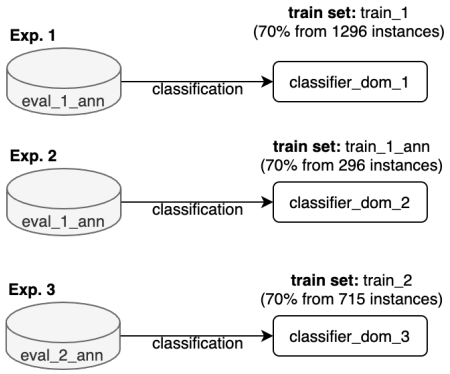

Figura 3.4: Descripción esquemática de los experimentos realizados sobre los modelos y dataset descritos en la tabla 4.2 y la figura 3.3. El clasificador *classifier_dom_3* es considerado la configuración final del Clasificador de Dominio.

Experimento 1

Queremos comprobar los efectos de combinar nuestro dataset anotado manualmente, *train_1_ann* a modo de semilla, sobre una base de 1000 instancias clasificadas por el Clasificador Genérico, *train_1_dev* en un dataset de 1296 instancias, *train_1* para construir un clasificador utilizando las *features* de la tabla 3.4.

Nuestra hipótesis consiste en que una semilla de este tipo (el 23 % del dataset), puede ser suficiente para generar una señal de clasificación de dominio con unos resultados aceptables.

Entrenamos el clasificador *classifier_dom_1* en una configuración similar a la utilizada en el Clasificador Genérico: Regresión Logística sobre *SGD*, con una división train-test del 70 %-30 %. Obtenemos unos resultados de extremo *over-fitting*, $F1$ =.99 sobre el set de test, probablemente por haber realizado nuestras observaciones utilizando un dataset tan limitado. Para tener una mejor idea del *performance* de nuestro modelo, lo evaluamos sobre un dataset manualmente anotado, *eval_1_ann* para el que obtenemos un $F1$ =.58, unos resultados aún muy por debajo de lo esperado, teniendo en cuenta que para este caso el *baseline* no-informativo es del 67 %.

Consideramos que este dataset y esta configuración no ha corroborado nuestra hipótesis para este experimento.

Experimento 2

Tras los resultados insuficientes obtenidos en el experimento anterior, nuestra siguiente hipótesis es que el dataset *train_1_dev* introduce ruido en la clasificación y que los resultados mejorarán al retirarlo del set de entrenamiento, entrenando sólo con aquellas instancias manualmente anotadas, el dataset *train_1_ann*. Entrenando el modelo *classifier_dom_2* en las mismas condiciones que en el experimento anterior y evaluando sobre el mismo dataset *eval_1_ann* se obtienen unos resultados muy similares, del $F1$ =.57, pero mejorando ligeramente la precisión de la clase *Argument*.

De aquí podemos concluir que el dataset *train_1_dev* no está aportando nada a la clasificación y que podemos retirarlo del set de entrenamiento, corroborando parcialmente nuestra hipótesis.

Experimento 3

Nuestra siguiente hipótesis es que los resultados mejorarán, utilizando sólamente instancias manualmente anotadas, con un dataset de mayor tamaño, *train_2*, la combinación de todas instancias manualmente anotadas que hemos usado hasta ahora. Para su evaluación, extraeremos otro dataset, *eval_2_ann*, el cual anotaremos manualmente de igual manera.

Entrenando el modelo *classifier_dom_3* en las mismas condiciones que en el experimento anterior y evaluando sobre el mismo dataset *eval_2_ann* se obtienen unos resultados considerablemente mejores de $F1$ =.69, que superan el *baseline* no-informativo para este caso, del 67 %. Por tanto damos nuestra hipótesis por válida para este experimento.

Conclusiones

Concluimos que los resultados obtenidos en esta última configuración: modelo lineal de aprendizaje máquina basado en optimizador *Stochastic Gradient Descent* de *Regresión Logística*, entrenado sobre un 70 % aleatoriamente seleccionado, conservando proporciones de clases, del dataset *train_2* son aceptables para nuestros propósitos en este componente, pero no deben ser considerado como clasificador único en el proceso.

A pesar de prácticamente igualar el $F1$ obtenido en el Clasificador Genérico (.69 frente a .73), no podemos apoyar el resto del proceso sólamente en la clasificación obtenida mediante este Clasificador de Dominio, debido al sesgo introducido al desarrollar todo el proceso sobre una muestra limitada de datos y realizar su evaluación sobre otra muestra también reducida y a una precisión menor en la detección de la clase *Argument* respecto al Clasificador Genérico.

Una alta precisión en la detección de la clase *Argument* es clave para el éxito del resto del *pipeline*, incluso si esto significa detectar un menor número de argumentos. A partir de los resultados obtenidos y nuestras observaciones, ambos clasificadores, Genérico y de Dominio parecen proporcionar visiones complementarias de la argumentación, capturando en un caso los elementos argumentativos de carácter general y en otro caso otros más específicos.

Esto nos hace plantear un sistema que combine dichas señales y otras que puedan resultar de interés para construir una señal final de mayor calidad.

3.2.2.3. Señal de Síntomas

Según diversos trabajos (Carrillo-de Albornoz et al., 2019) centrados en el procesamiento de lenguaje natural para el ámbito médico, la presencia de descripciones de síntomas y efectos secundarios son un componente muy relevante a la hora de detectar argumentación en textos de este tipo y este ámbito. Este comportamiento ha sido igualmente confirmado por nuestras observaciones, derivadas del análisis cualitativo, mencionado en la sección anterior.

Estas características ya han sido implícitamente incluídas en las *features* descritas en la tabla 3.4, en forma de número normalizado de síntomas/efectos secundarios presentes en la oración. No obstante, parece conveniente reforzar la aportación de la presencia de estos síntomas a nuestros argumentos en el dominio del ámbito médico. Por tanto, se incluye esta señal binaria de presencia de síntomas en la oración como una tercera señal relevante para mejorar la precisión con la que identificamos la clase *Argument*, en combinación con las dos señales ya descritas, provistas por el Clasificador Genérico y Clasificador de Dominio.

3.2.2.4. Voto Mayoritario

Como podemos ver en la figura 3.2, combinamos las tres señales:

- Clasificador Genérico
- Clasificador de Dominio
- Señal de Síntomas

Y extraemos una señal de salida como voto mayoritario de las tres. Esta será nuestra clasificación definitiva de argumentación a nivel de oración para el resto del proceso, obteniendo unos resultados de $F1$ =0.68 sobre el set de evaluación anteriormente descrito *eval_2_ann* (ver tabla 4.2), con una clase mayoritaria del .67. Sin embargo, estamos consiguiendo una precisión muy elevada del .92, un recall por encima del recall obtenido por dos de las señales anteriores. [8]

[8] Resultados ampliados en la sección de Resultados, tabla 4.6

Con esta señal final, podemos asumir que pueden conseguirse resultados mínimos viables para poder construir con éxito el resto del *pipeline*.

3.2.3. Clustering de Argumentos

Una vez que tenemos un set suficiente de instancias argumentativas, podemos proceder al paso de agrupamiento o clustering de las mismas. Identificaremos posteriormente un clúster con una *sub-temática* o argumento del tema de controversia que estamos considerando.

Tras una primera exploración de un clustering *clásico*, a través del uso de un algoritmo partitivo como *K-Means* utilizando *tf-idf* como *features*, encontramos varios inconvenientes:

- No podemos conocer el número de *clusters a priori*. De hecho esta es una de las informaciones que nos interesan.
- Si realizamos varias ejecuciones para diferentes números de *clusters* K, no se llega a apreciar convergencia para ningún número distinto al número de elementos individuales. Lo cual no nos aporta información, al ser una solución trivial.
- Las features de frecuencia de términos *tf-idf* no nos proporcionan el aspecto semántico que buscamos en un clúster.

Por estas razones, exploramos la vía del clustering aglomerativo jerárquico, presentado en la sección 2.4

Utilizamos la representación vectorial de las instancias, a través de los vectores presentados en (Baroni, Dinu, y Kruszewski, 2014), ya extraídos anteriormente como *features* y calculamos su distancia de coseno por parejas para construir una matriz de similitud. Mediante el uso de los *word-embeddings* estamos garantizando que se tenga en cuenta la componente semántica que realmente nos interesa, para construir temas mínimamente coherentes en los *clusters*.

A partir de su matriz de similitud, realizamos el clustering aglomerativo jerárquico a través del método de Ward (Ward Jr, 1963), basado en buscar una varianza mínima a la hora de combinar dos *clusters*.

Mediante este método, obtendremos un árbol de jerarquías de similitud de los diferentes *clusters*, un dendograma como el mostrado en la figura 3.5.

Sobre esta estructura de árbol, aún tenemos que decidir con qué nivel de dicho árbol nos quedamos. Si elegimos una alta similitud, encontraremos *clusters* muy grandes con temáticas mezcladas y si elegimos una baja similitud, obtendremos muchos clúster de poco tamaño. Realizando varias ejecuciones y llevando a cabo un análisis cualitativo del contenido de los *clusters* y de su tamaño, hemos elegido empíricamente un valor del 30 % del máximo de similitud, para obtener un número de *clusters* de tamaño y coherencia aceptables.

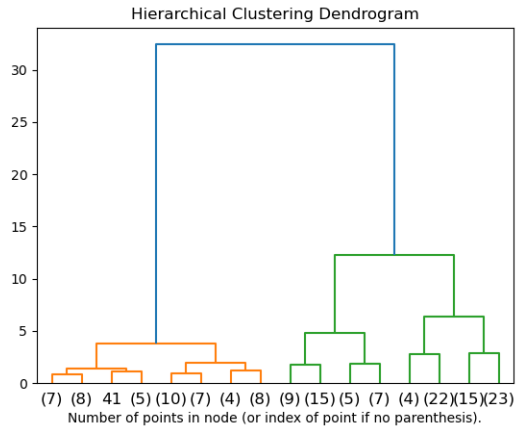

Figura 3.5: Ejemplo de dendograma obtenido a partir de la clusterización aglomerativa jerarquizada.

Este umbral puede estimarse con técnicas más avanzadas, pero no se quedan fuera de los márgenes de los objetivos para este trabajo. Respecto al clustering, necesitamos solamente una funcionalidad mínima viable para poder ejecutar y evaluar el proceso en su conjunto.

3.2.4. Clasificación de Polaridad

Aplicamos una de las técnicas del estado del arte, presentadas en la sección 2.5, a través de la construcción de un modelo supervisado de clasificación basado en redes neuronales de tipo LSTM (Tai, Socher, y Manning, 2015).

Para la construcción del modelo utilizaremos el *word-embedding* (Baroni, Dinu, y Kruszewski, 2014), ya utilizado anteriormente durante la extracción de *features*, para representar generar una representación vectorial de nuestras oraciones. Necesitamos, además, una anotación de sentimiento a nivel de oración que nos permita entrenar el clasificador sobre las categorías ("positivo" vs. "negativo").

Al carecer de recursos y no constituir este paso un punto crítico ni una tarea sin resolución, en el grado que necesitamos en este *pipeline*, se ha decidido usar una variable de *rating* que forma parte del dataset de *Drug Review Dataset*, para entrenar y evaluar el desempeño de este componente.

Esta variable *rating* tiene un valor de 1 a 10 que mide la satisfacción del usuario respecto al medicamento en su caso de uso, y esta asociada a nivel de comentario. Esto quiere decir, que primeramente realizaremos una extrapolación de dicho *rating* a las diferentes oraciones que componen el comentario, bajo la aproximación de que un comentario negativo (positivo) implicará elementos negativos (positivos) en todas las oraciones que lo componen. A continuación, se ha discretizado los valores

de dicho *rating* en "positivo" para un valor mayor que 5, y "negativo" en caso contrario, en base a un análisis manual de los textos.

Entrenando una red *LSTM* en su configuración más básica obtenemos un *accuracy del 77.8 %*, lo cual nos parece adecuado para nuestro propósito.

El resultado final será la asignación de una clase a cada oración ("positivo" vs. "negativo"), que serán usadas para el cómputo de una polaridad neta de un clúster, a través de los sentimientos de los elementos que lo forman.

3.2.5. Estimación de Controversia

Ahora que tenemos un sistema con pasos definidos, podemos rescatar la definición 1 para desarrollar una definición más técnica para la estimación del valor de controversia. Tras procesar un tema a través del *pipeline* descrito, obtendremos un escenario en el que tenemos diversos *clusters* de diferentes tamaños con polaridades internas. Podemos utilizar estas propiedades para realizar una estimación de la controversia:

Definición 2. Dado un tema T compuesto por el conjunto de $\{d_1, ..., d_m\}$ documentos (*comentarios*), segmentable en el conjunto de elementos textuales únicos (*oraciones*)$S = \{s_1, s_2, ..., s_p\}$, de los cuales extraemos aquellos que manifiesten argumentación en el subconjunto $A \subseteq S$.
Se dirá que es T será controvertido si es posible la formación de N clusters $a_i \in A \mid i = 1, ..., N$ de tamaños relevantes, y comparables de polaridades netas $pol(a_i)$ opuestas y pudiendo estimarse su controversia como:

$$C = \frac{1}{N} \sum_{c=1}^{N} s(a_c) \cdot pol(a_c) = \frac{1}{N} \sum_{c=1}^{N} s(a_c) \cdot (p_c - n_c) \tag{3.1}$$

Siendo p_i y n_i el número de argumentos clasificados como positivos y negativos, respectivamente y $s(a_i)$ el tamaño del clúster a_i.

Lo cual puede interpretarse como la media de las polaridades netas de los clúster encontrados en un tema, pesados por el tamaño de estos.

El resultado de la estimación, será un número real, cuyo signo representa la polaridad global del tema ("positiva" vs. "negativa") y que será más controvertido cuanto menor sea $\mid C \mid$.

Resultará más sencillo distinguir los temas que no son en absoluto controvertidos (muestran consenso) que intentar ordenar en un *ranking* de manera cuantitativa aquellos que sí lo son, dado el nivel de sutileza.

Este aspecto será clave para poder evaluar el proceso entero. El evaluador debe tener en cuenta, que todo lo que sea claramente "no-controvertido" debe etiquetarse como tal, y en caso contrario será "controvertido".

A pesar de que estemos proporcionando una estimación numérica, dicha estimación será orientativa en el marco de este trabajo. Además, se realizará un análisis

cualitativo de los resultado de evaluación final, ya que serían necesarios unos recursos de anotación y marco de tiempo más amplio para poder realizar un análisis cuantitativo más exhaustivo.

Cabe mencionar que el objetivo de nuestro sistema es el de detectar la controversia, utilizando una cuantificación orientativa, pero definitivamente su potencial subyace en mayor medida en los aspectos cualitativos de dicha controversia, más que en la capacidad de proporcionar una cuantificación exacta. En ese sentido, nuestra definición ha aportado resultados prometedores.

Capítulo 4

Evaluación y Discusión

En este capítulo analizaremos, por una parte de los componentes implicados en el paso de *Detección de Argumentos* y por otra parte una evaluación del *pipeline* al completo tras realizar una ejecución sobre una muestra de nuestro corpus objetivo de fármacos *Drug Review Dataset*.

La evaluación de la *Detección de Argumentos* se realizará para los tres subcomponentes por separado (Clasificador Genérico, Clasificador de Dominio y Voto Mayoritario) y se presentarán los resultados obtenidos y decisiones tomadas. La evaluación de la clasificación sobre el corpus de fármacos requerirá de varias iteraciones de anotación manual para su comparación con los resultados proporcionados por los diferentes modelos.

La evaluación del *pipeline* al completo también requerirá de una serie de anotaciones manuales con las que comparar nuestros resultados. La evaluación manual intentará proporcionar un etiquetado binario, que deberá ser comparado en forma de análisis cualitativo con los resultados obtenidos por el sistema, así como su discusión.

El componente de clusterización ha sido evaluado cualitativamente mediante la comprobación manual de coherencia temática en un conjunto de 20 *clusters* seleccionados aleatoriamente de entre varios casos de uso de fármacos y se ha considerado aceptable, dentro del marco de este trabajo, para realizar dicha funcionalidad en el *pipeline*.

De manera similar, el paso de *Sentiment Analysis*, no ha sido exhaustivamente evaluado y optimizado. Se ha utilizado una configuración básica del estado del arte, sobre un set de 60k instancias, para el que se ha obtenido un *accuracy* del 78 %, utilizando la configuración descrita en la sección 3.2.4 para clasificar nuestras instancias como “positivas” o “negativas”. Este *setup* se ha considerado aceptable para la función que cumple en nuestro *pipeline*.

4.1. Detección de Argumentos

Para el desarrollo de nuestro componente de detección de argumentos, se han llevado a cabo diferentes experimentos, como se describe en la sección 3.2.2, para sus tres sub-componentes: Clasificador Genérico, Clasificador de Dominio y Voto Mayoritario. Presentaremos en los siguientes apartados, los resultados obtenidos, siguiendo su orden de ejecución, así como la interpretación y conclusiones sobre dichos resultados.

Como un esquema previo, en analogía a la estructura de la sección 3.2.2, encontramos:

1. **Clasificador Genérico**: nuestra hipótesis consiste en que pueden obtenerse resultados satisfactorios de clasificación de argumentos utilizando las *features* (tablas 3.1 y 3.2) sobre un corpus de amplio dominio *Cross-Domain Sentence Argument Mining* para identificar argumentos en nuestro corpus de dominio *Drug Review Dataset*, que mejoren nuestro *baseline* no-informativo. Distinguimos dos etapas:

   - Validación del uso de las *features* mencionadas para construir un clasificador binario de argumentación, basado en un modelo de aprendizaje máquina, entrenado y evaluado con el corpus *Cross-Domain Sentence Argument Mining*, con resultados que mejoren nuestro *baseline* no-informativo y sean similares a los del estado del arte para esta tarea (Stab et al., 2018).
   - Validación del uso del clasificador construido en la etapa anterior en su configuración óptima para la detección de argumentos en el corpus *Drug Review Dataset*, evaluado manualmente mediante un análisis cualitativo de la señal de clasificación.

2. **Clasificador de Dominio**: nuestra hipótesis consiste en que pueden obtenerse resultados satisfactorios de clasificación de argumentos para el corpus *Drug Review Dataset* mediante la extracción de *features* de dominio y el uso de una semilla extraída del corpus *Drug Review Dataset* (Gräßer et al., 2018), anotada manualmente, que mejoren nuestro *baseline* no-informativo. Se distinguen tres experimentos[1]:

   - Experimento 1: la hipótesis consiste en que pueden obtenerse resultados satisfactorios, por encima de nuestro *baseline* no-informativo y de los resultados obtenidos por el Clasificador Genérico, combinando un dataset anotado manualmente y la salida del Clasificador Genérico, en su configuración final, extraídos ambos de *Drug Dataset Review*, para

[1] descripción de los datasets en la tabla 4.2

entrenar un clasificador binario basado en un modelo de aprendizaje máquina.

- Experimento 2: la hipótesis consiste en probar si el Clasificador Genérico introduce ruido en el Clasificador de Dominio. Para ello, realizaremos de nuevo el experimento anterior, utilizando como set de entrenamiento sólamente aquellas instancias anotadas manualmente.
- Experimento 3: la hipótesis consiste en probar que los resultados de los experimentos 1 y 2 mejorarán utilizando un dataset constituido únicamente por instancias de *Drug Dataset Review* anotadas manualmente y en mayor número.

3. **Voto Mayoritario**: nuestra hipótesis consiste en que una combinación de señales que incluyan las obtenidas a partir del Clasificador Genérico y del Clasificador de Dominio y un señal adicional de dominio puede resultar en una clasificación de mayor calidad que cualquiera de los componentes por separado.

Métricas y Criterios de Evaluación

Los elementos a evaluar en este componente son modelos de clasificación binaria, y como tales, utilizaremos como métricas de evaluación las medidas de Precisión, *Recall* y F1, éste último como medida más representativa de los resultados, que a su vez es la media armónica de las dos anteriores. Usaremos las siguientes definiciones:

$$Precision = \frac{T_P}{T_P + F_P} \tag{4.1}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{4.2}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.3}$$

donde T_P, F_P y F_N corresponden respectivamente a positivos verdaderos, falsos positivos y falsos negativos. En igualdad de condiciones para dos resultados con misma $F1$, consideramos en nuestro caso más relevante una precisión alta, que un *recall* alto. De esta manera, estamos favoreciendo el obtener casos claros de argumentación, aunque no consigamos identificar demasiados, a cambio de no introducir ruido innecesario.

Nuestro *baseline*, con el que compararemos los resultados de clasificación obtenidos en los diferentes experimentos, será el de un *baseline* no-informativo, marcado por la proporción de la clase mayoritaria en el dataset.

Nuestro sistema tiene como elemento fundamental de base la detección de argumentos, y debemos garantizar unos resultados suficientes para poder desarrollar

con ciertas garantías los siguientes componentes y no arrastrar una alta tasa de error desde el primer paso del *pipeline*.

4.1.1. Clasificador Genérico

Nuestro objetivo consiste, en primer lugar, en demostrar que podemos utilizar las *features* (tablas 3.1 y 3.2)(Stab y Gurevych, 2017), diseñadas para la clasificación de sub-componentes de argumentación en ensayos persuasivos, para construir un clasificador de argumentos genérico, usando como set de entrenamiento el corpus *Cross-Domain Sentential Argument Mining* (Stab et al., 2018).[2]

Para ello, se han considerado 19,443 oraciones[3], pertenecientes al corpus de 25,000 oraciones, anotadas como (*argument-pro*, *argument-against*, *no-argument*) presentado en (Stab et al., 2018). Las anotaciones de *argument-pro* y *argument-against* se han unificado en una sola clase, para realizar una clasificación binaria: *argument* vs. *no-argument*.

	No-Argument			*Argument*			***All***
	P	*R*	*F1*	*P*	*R*	*F1*	*Avg. F1*
SVM + all features	.79	**.90**	.77	**.78**	.44	.56	.66
SVM + all features except word-embeddings	.62	.94	.75	.80	.28	.41	.58
ModHuber + all features	**.85**	.44	.58	.56	**.90**	.69	.63
Huber + all features	.79	.69	.74	.66	.77	**.71**	.72†
LogReg + all features	.74	.81	**.78**	.73	.64	.68	**.73**†

Tabla 4.1: Resultados relevantes de clasificación sobre el set de evaluación de argumentos en diferentes modelos lineales sobre el optimizador SGD: Support Vector Machine (SVM), Huber Regression (HuberReg), Modified Huber Regression (ModHubReg), Logistic Regression (LogReg) († = mejora sobre el *baseline* y sobre los resultados del estado del arte para este corpus)

Se ha dividido el corpus aleatoriamente en dos sets 70 % y 30 % para entrenamiento y test, respectivamente. Se ha realizado para la evaluación un *cross-validation* de 5x2.

El set de test consiste, por tanto, en 5833 instancias, cuya clase mayoritaria es *no-argument*, que constituye un 56 % del total de instancias, por lo que nuestro *baseline* no-informativo de referencia será $F1 = .56$.

Se han seleccionado varios modelos mencionados en la sección 2.3, para los que se ha realizado una hiperparametrización, y se han usado junto con el optimizador *Stochastic Gradient Descent (SGD)*[4] para entrenar sobre las 7849 *features* presentadas en 3.1.

[2] analizado en la sección 2.3.1

[3] por un fallo técnico en el proceso de extracción, no se ha utilizado el total del corpus

[4] utilizado normalmente para una rápida convergencia de las funciones de aprendizaje.(Bach y Moulines, 2011)

Los resultados más representativos de dicho proceso se muestran en la tabla 4.1, alcanzándose el mejor resultado para el modelo de Regresión Logística con $F1$ =**.73**, mejorando el *baseline* aleatorio de .56, y los resultados del estado del arte presentados en (Stab et al., 2018) para clasifiación binaria sobre este corpus ($F1$ =.67). Adicionalmente al mejor $F1$ medio (.73), se ha considerado que esta configuración es la que proporciona un mejor equilibrio a la precisión y *recall* de la clase *Argument*, sin impactar al eficacia sobre la otra clase y sobreponiendo precisión a *recall*.

Como detalle de interés, podemos ver como la inclusión o exclusión de *word-embeddings* como *feature* puede impactar en gran medida al *recall* de la clase *Argument*, como puede verse en los dos resultados mostrados del modelo *SVM*.

Consideramos que estos resultados satisfacen los requerimientos de nuestro primer objetivo: las *features* seleccionadas son válidas sobre el corpus seleccionado *Cross-Domain Sentential Argument-Mining* y por tanto, consideraremos la configuración ***LogReg + all features*** como nuestra **configuración final** para el Clasificador Genérico.

Nuestro segundo objetivo es el de evaluar la validez de este modelo sobre nuestro dataset de dominio *Drug Review Dataset*.

Consideramos el dataset *dev* (ver tabla 4.2), consistente en 5526 instancias de *Drug Review Dataset*, y las clasificamos mediante la configuración final del Clasificador Genérico, para su análisis cualitativo.

Exp.	Dataset	Instancias	Descripción
-	*pharma_dev*	68718	extraído aleatoriamente de *Drug Review Dataset*.
-	*dev*	5526	extraído aleatoriamente a partir de *pharma_dev*.
1	*train_1_dev*	1000	extraído aleatoriamente a partir de *dev*, clasificado por el Clasificador Genérico en su configuración final.
1, 2	*train_1_ann*	296	extraído aleatoriamente a partir de *dev* y anotado manualmente. Usado para crear *train_1* y para entrenar *classifier_dom_2*
1	*train_1*	1296	combinación de *train_1_dev* y *train_1_ann*, usado para entrenar *classifier_dom_2*.
2	*eval_1_ann*	419	extraído aleatoriamente de *dev* y anotado manualmente. Usado para evaluar *classifier_dom_1* y *classifier_dom_2*.
3	*train_2*	715	combinación de *train_1_ann* y *eval_2_ann*. Usado para entrenar *classifier_dom_3*.
3	*eval_2_ann*	215	extraído aleatoriamente de *dev* y anotado manualmente. Usado para evaluar *classifier_dom_3*.

Tabla 4.2: Datasets utilizados para la construcción del Clasificado de Dominio. El proceso de extracción de dichos datasets puede encontrarse en la sección 3.2.2.2.

En dicho análisis, se observa que la detección de la clase de mayor interés, *Argument*, es considerablemente insuficiente, debido al alto número de instancias que en el dominio de *Drug Review Dataset* se consideran argumentos y no son clasificadas como tal, en lo que parece ser un escenario de alta precisión pero muy bajo *recall*.

Los resultados de clasificación muestra un 85.76 % de instancias clasificadas como *No-Argument* y un 14.24 % de instancias de la clase *Argument*, muy por debajo de lo que cabría esperarse en este dataset.

Una porción de casos de argumentación de ámbito general son detectados, pero en el ámbito de este corpus concreto se han observado una serie de *features* que concretan los mecanismos generales de argumentación.

Este resultado no satisface nuestro segundo objetivo, lo cual nos hace plantear el desarrollo de un segundo clasificador de dominio, partiendo de la señal de clasificación del Clasificador Genérico.

4.1.2. Clasificador de Dominio

Nuestro objetivo es el de probar que pueden obtenerse resultados satisfactorios de clasificación de argumentos para el corpus *Drug Review Dataset*, mediante un clasificador basado en *features* de dominio (ver tabla 3.4) que nos permita mejorar los resultados del Clasificador Genérico, detectando todos aquellos casos, en especial de la clase *Argument*, no detectados por el Clasificador Genérico.

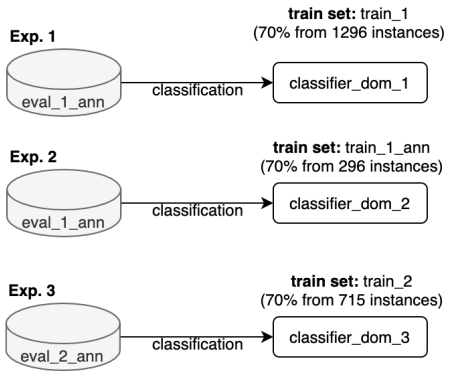

Figura 4.1: Descripción esquemática de los experimentos realizados sobre los modelos y dataset descritos en la tabla 4.2 y la figura 3.3. El clasificador *classifier_dom_3* es considerado la configuración final del Clasificador de Dominio.

Se desarrollan para ello tres experimentos (descritos en la sección 3.2.2.2), utilizando los datasets descritos en la tabla 4.2, que podemos apreciar de manera esquemática en la figura 4.1.

4.1.2.1. Experimento 1

Se pretende validar la combinación de la salida del Clasificador Genérico y una semilla de *Drug Review Dataset* anotada manualmente para entrenar un clasificador de dominio con el que conseguir resultado más satisfactorios sobre *Drug Review Dataset* que los obtenidos mediante el Clasificador Genérico.

Experimento 1.1

Utilizamos el dataset *train_1*, compuesto por el dataset *train_1_dev*, clasificado por el Clasificador Genérico y el dataset anotado *train_1_ann*, anotado manualmente, para mediante el uso de las *features* de dominio (tabla 3.4), entrenar un modelo lineal de aprendizaje máquina de tipo Regresión Logística con optimizador *Stochastic Gradient Descent* y una división train/test de 70 %/30 % (*classifier_dom_1*). Obtenemos un $F1$ =.99 medio sobre el set de test, lo cual consideramos una situación de extremo *over-fitting*.

Experimento 1.2

Realizamos a continuación una evaluación del modelo entrenado anterior, entrenado con *train_1* usando el dataset *eval_1_ann*, anotado manualmente, como set de evaluación. Obtenemos los resultados mostrados en la tabla 4.3, con un $F1$ = ,58 medio, muy por debajo de nuestro *baseline* no-informativo del .67 sobre la clase mayoritaria.

	P	R	F1	Support
No-Argument	.72	.94	.81	283
Argument	.65	.24	.35	136

Tabla 4.3: Resultados del Experimento 1. Esquema en Fig. 4.1.

Consideramos como inválida nuestra hipótesis, descartando esta configuración.

4.1.2.2. Experimento 2

Nuestra hipótesis parte de que el dataset *train_1_dev* es relevante para la clasificación. Para ello, queremos comprobar si el dataset *train_1_dev*, la salida del Clasificador Genérico, está generando ruido en la clasificación y si entrenando un clasificador en la misma configuración, pero utilizando sólamente el sub-dataset anotado manualmente *train_1_ann* como entrenamiento, podemos obtener mejores o iguales resultados.

Aplicamos de nuevo el mismo procedimiento, mediante el uso de las *features* de dominio (tabla 3.4), entrenamos un modelo lineal de aprendizaje máquina de tipo

Regresión Logística con optimizador *Stochastic Gradient Descent* y una división train/test de 70 %/30 % (*classifier_dom_2*).

Evaluando sobre el mismo set de evaluación que en el experimento anterior, *eval_1_ann*, obtenemos los resultados mostrados en la tabla 4.4, muy parecidos a los anteriores, con un $F1$ =.57 medio, lo que demuestra que el dataset *train_1_dev* no está aportando a la clasificación, invalidando nuestra hipótesis de que dicho dataset era relevante.

	P	R	F1	Support
No-Argument	.71	.95	.82	283
Argument	.67	.21	.31	136

Tabla 4.4: Resultados del Experimento 2. Esquema en Fig. 4.1.

Los resultados siguen sin ser satisfactorios, estando aún por debajo del *baseline* no-informativo de .67.

4.1.2.3. Experimento 3

Partiendo de los resultados de los Experimentos 1 y 2, queremos entrenar un nuevo clasificador utilizando únicamente instancias que hayan sido manualmente anotadas. Para ello combinamos todas las instancias anotadas que hemos utilizado hasta el momento, *train_1_ann* y *eval_1_ann*, combinadas como *train_2* para entrenar el clasificador *classifier_dom_3* con un modelo lineal de aprendizaje máquina de Regresión Logística sobre optimizador *Stochastic Gradient Descent*.

Para su evaluación, realizamos una nueva extracción y anotación manual para crear el dataset *eval_2_ann*, para el cual obtenemos los resultados mostrados en la tabla 4.5, con un $F1$=**.69** medio, por encima de nuestro *baseline* no-informativo del .67.

	P	R	F1	Support
No-Argument	.78	.86	.82	145
Argument	.72	.50	.56	70

Tabla 4.5: Resultados del Experimento 3. Esquema en Fig. 4.1.

Validamos así nuestra hipótesis, de que podemos conseguir resultados satisfactorios usando datasets formados únicamente por instancias anotadas manualmente, usando las *features* de dominio, presentadas en la tabla 3.4.

Consideraremos, por tanto, la configuración definida por el modelo *classifier_dom_3* como **configuración final** para el Clasificador de Dominio.

Sin embargo, consideramos que este clasifiador puede resultar insuficiente como único clasificador de argumentos en el sistema. La baja $F1$ para la clase *Argument*, el desarrollo de *features* basado en extracciones parciales del corpus, así como el

entrenamiento utilizado basado en extracciones parciales puede resultar en una clasificación muy sesgada para un conjunto de datos.

Esto nos hace plantear una solución que combine las señales de ambos clasificadores y alguna señal adicional de dominio para construir la señal final de clasificación.

4.1.3. Voto Mayoritario

Nuestra hipótesis para este bloque consiste en que la combinación de las señales de clasificación del Clasificador Genérico y del Clasficador de Dominio, junto a alguna señal adicional de dominio, aportará unos resultados satisfactorios de clasificación.

En este caso estamos realizando la combinación de los clasificadores anteriores, más una señal binaria que indica la presencia o no de síntomas en la oración (sección 3.2.2.4). El resultado final de la clasificación binaria será el voto mayoritario de las tres señales mencionadas y será el resultado que consideremos como válido para los siguientes pasos del *pipeline*. Al ser otro clasificador binario más, utilizamos por tanto la misma métrica que en los anteriores.

Se han seleccionado como test de evaluación, el dataset utilizado anteriormente para evaluar la configuración final del Clasificador de Dominio: *eval_2_ann*.

Evaluando las tres señales por separado (estando Clasificador Genérico y Clasificador de Dominio en configuración final) y su combinación mediante el voto mayoritario, obtenemos los resultados mostrados en la tabla 4.6.

	No-Argument			*Argument*			***All***
	P	*R*	*F1*	*P*	*R*	*F1*	*Avg. F1*
Generic Classifier	.73	.95	.82	.72	.26	.38	.60
Domain Classifier	**.78**	.86	.82	.62	**.50**	**.56**	**.69**†
Symptom Signal	.73	.93	.82	.68	.30	.42	.62
Majority Vote	.76	**.99**	**.86**	**.92**	.34	.50	.68†

Tabla 4.6: Resultados de clasificación sobre el set de 215 instancias de argumentos en diferentes clasificadores: Genérico (LogReg de la sección 4.1.1, Dominio (LogReg con *features* de dominio) y Señal Binaria de Síntomas), († = mejora sobre el *baseline*).

Con el voto mayoritario, obtenemos un $F1 = .68$ medio, con el que conseguimos mejorar ligeramente el *baseline*, sacrificamos el *recall* (.34) de la clase *Argument*, a cambio de alcanzar una precisión del .92, al mismo tiempo que mantenemos buenos resultados para la clase *No-Argument*. Confirmamos así nuestra hipótesis para este bloque.

Nuestro objetivo en este paso no es construir un clasificador perfecto para ambas clases, sino ser capaz de extraer tantos elementos de la clase *Argument* como sea posible, con la máxima precisión.

Por estas razones y por una potencial falta de generalización, seleccionamos el modelo de *Voto Mayoritario* frente al Clasificador de Dominio, que aporta una mejor $F1$=.69, ya que este ha sido entrenado sobre una extracción limitada y potencialmente sesgada del corpus *Drug Review Dataset*.

4.2. Estimación de Controversia

4.2.1. Experimento

El objetivo de este experimento es comprobar que nuestro marco para la identificación de controversia, consistente en nuestra propuesta de definición de controversia y nuestra propuesta de *pipeline* para su detección, nos permite obtener resultados satisfactorios en su aplicación al caso de estudio de este trabajo, el dataset *Drug Review Dataset*.

Para ello, realizaremos, por un lado, una extracción aleatoria de 23,805 instancias (oraciones) a partir de *Drug Review Dataset*, preservando la unidad de los comentarios de usuarios de las que proceden.

Seguidamente, realizaremos una ejecución de nuestro *pipeline*, en su configuración final descrita en el capítulo de Sistema (3), sobre esta extracción, obteniendo los consiguientes resultados para la estimación de controversia para los diferentes casos de uso[5] presentes en el dataset.

Para evaluar dichos resultados, seleccionaremos aleatoriamente 10 casos de uso de los considerados para la estimación, sin repetir el fármaco o la enfermedad que componen el caso. Necesitaremos realizar un proceso de evaluación que determine si se tratan de casos *controvertidos* o *no-controvertidos*

Para dicha evaluación, se realiza un proceso de anotación con tres anotadores humanos independientes, partiendo de la definición propuesta de controversia como criterio de anotación, que rercordemos, consiste en: *Dado un tema de discusión determinado, y un conjunto de elementos de contenido textual generados por una comunidad de individuos en relación a dicho tema, se considerará que tal tema posee un grado medible de controversia respecto a tal comunidad, si:*

1. *Puede extraerse un conjunto suficiente de unidades textuales con indicios de* ***argumentación****, a los que llamaremos argumentos.*
2. *Dichos argumentos pueden agruparse en un conjunto de* ***grupos****, que constituirán las temáticas sobre las que se apoyan las posiciones.*
3. *Dichos grupos poseen una* ***polaridad*** *que puede ser evaluada.*
4. *Grupos de tamaño relevante y comparable entre sí manifiestan* ***polaridades enfrentadas****.*

[5] recordemos que un caso de uso consiste en la combinación fármaco+enfermedad

Finalmente, se realizará un análisis cualitativo de los resultados obtenidos mediante nuestro *pipeline* y el voto mayoritario de los anotadores, para extraer en qué casos nuestro sistema realiza estimaciones correctas de controversia y qué características están presentes en aquellos casos en los que los anotadores y los resultados obtenidos difieren.

4.2.1.1. Métricas de evaluación

Durante el proceso de evaluación de los resultados, se utilizarán las siguientes métricas para determinar el grado de acuerdo entre anotadores:

- **Acuerdo**: media del acuerdo entre evaluatores, marcando +1 para acuerdo entre dos evaluadores y 0 para desacuerdo.
- **Kappa de Cohen** (Carletta, 1996):

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

  donde $Pr(a)$ es el acuerdo observado relativo entre los observadores, y $Pr(e)$ es la probabilidad hipotética de acuerdo por azar, utilizando los datos observados para calcular las probabilidades de que cada observador clasifique aleatoriamente cada categoría.

Los valores de κ se interpretarán según (Landis y Koch, 1977), cuyo criterio se muestra en la tabla 4.7.

grado de acuerdo	κ
leve	0 - 0.20
aceptable	0.21 - 0.40
moderado	0.41 - 0.60
sustancial	0.61 - 0.80
casi perfecto	0.81 - 1

Tabla 4.7: Criterios para κ de Cohen

Para la evaluación final de resultados, es decir, la comparación entre los resultados obtenidos mediante nuestro sistema y la anotación realizada por los anotadores, no se han utilizado métricas cuantitativas.

Dado el carácter sutil de la evaluación de controversia, los limitados recursos disponibles para realizar una anotación más compleja y la dificultad de establecer un *threshold*, que nos proporcionara conclusiones más definitivas, así como la ausencia de un marco de evaluación común en el estado del arte, se ha procedido a realizar un análisis cualitativo de los resultados. Dicho análisis nos permitirá entender cómo ha sido el funcionamiento de nuestro *pipeline*, analizando las características presentes en los diferentes grados de coincidencia entre sistema y anotadores e identificando sus fortalezas y debilidades.

4.2.2. Ejecución del *Pipeline*

Para la ejecución del *pipeline* sobre nuestra muestra de 23,805 instancias de *Drug Review Dataset*, se han configurado los componentes del *pipeline* (Fig. 3.2), utilizando las configuraciones óptimas descritas en esta sección para los modelos de detección de argumentación, que recordemos son:

- El Clasificador Genérico, consistente en un modelo lineal de aprendizaje máquina de tipo Regresión Logística sobre un optimizador *Stochastic Gradient Descent*, entrenado sobre el 70 % seleccionado aleatoriamente de las 19,443 instancias que componen dataset del corpus *Cross-Domain Sentential Argument Mining* (Stab et al., 2018) utilizando las *features* de las tablas 3.1 y 3.2.

- El Clasificador de Dominio, entrenado con el subconjunto *train_2* (ver tabla 4.2) del corpus *Drug Review Dataset*, sobre un modelo de Regresión Logística sobre un optimizador *Stochastic Gradient Descent* utilizando las *features* descritas en la tabla 3.4.

Entre los resultados obtenidos, hemos seleccionado 10 casos de uso aleatoriamente, sólamente teniendo en cuenta que no se repitan fármaco o enfermedad en el conjunto seleccionado.

Los casos de estudio seleccionados y los resultados de estimación de controversia obtenidos tras aplicar nuestro *pipeline* se presentan en la tabla 4.8.

id	Fármaco	Uso	C
1	adapalene/benzoyl peroxide	acné	-1.441
2	escitalopram	depresión	2.372
3	drospirenone/ethinyl estradiol	control de natalidad	-2.889
4	contrave	obesidad	3.023
5	lamotrigine	desorden bipolar	-3.221
6	clonazepam	ansiedad	3.331
7	vyvanse	TDAH	4.890
8	bisacodyl	estreñimiento	-5.636
9	chantix	dejar de fumar	-7.084
10	miconazole	candidosis vaginal	-22.026

Tabla 4.8: Casos seleccionados para la estimación de controversia. Un caso está compuesto por un fármaco en un uso particular (sintomatología, condición, enfermedad, etc.). El signo de la controversia indica la polaridad global del tema. La medida de controversia se encuentra ordenada de menor a mayor en valor absoluto para una mejor representación (recordemos que el máximo de controversia se dará para un escenario donde $C = 0$ y se dará algo o nada de controversia para valores mayores de $\mid C \mid$).

4.2.3. Anotación de los Casos de Uso

La evaluación manual de controversia se ha realizado con tres evaluadores independientes y consiste en las siguientes fases:

1. Se les ha presentado a los anotadores el criterio de anotación, consistente en nuestra definición de controversia (sección 3.1.1), tal como se ha descrito en la presentación del experimento (sección 4.2.1).
2. Cuando el proceso de anotación se aproxima al 20 % de su desarrollo, se realiza una puesta en común sobre el criterio de aplicación de la definición de cada uno de los anotadores al caso particular de *Drug Review Dataset*. En esta puesta en común no se intercambian detalles de anotación ni de casos concretos.
3. A partir de la puesta en común, se evidencia la dificultad, en varios casos, de determinar el límite a partir del cual puede considerarse un caso como controvertido. Para hacer esta distinción, cada anotador se basaba en criterios propios, como la identificación de patrones que hicieran compatibles las diferentes experiencias encontradas en los comentarios o en criterios ligados concretamente al ámbito médico, como la intensidad o gravedad de los efectos secundarios de un fármaco.
4. Para reducir dicha subjetividad, se propone una matización en el criterio de anotación, bajo la premisa de la definición, que facilite la tarea de anotación para los anotadores, además de estandarizar el criterio:

   > aquellos casos en los que se evidencie una ausencia de controversia, serán anotados como *no-controvertidos*, mientras que aquellos que presenten cierto nivel, aunque leve, de controversia se considerarán *controvertidos*.

5. Se reinicia el proceso de anotación utilizando la matización anterior del criterio de anotación, bajo la premisa de la definición.
6. Tras finalizar el proceso de anotación, se pide un breve *feedback* a los anotadores, recopilando las impresiones que la anotación de los diferentes casos les ha causado.

Los resultados obtenidos de dicha anotación se presentan en la tabla 4.9, junto al voto mayoritario de los anotadores:

Hemos añadido en la tabla 4.9 los valores para la controversia estimada por el sistema, para comparar con los resultados del voto mayoritario de los anotadores.

Aplicando las métricas de acuerdo definidas en la descripción del experimento (sección 4.2.1), obtenemos los valores presentados en la tabla 4.10.

Según los criterios presentados en la descripción de dichas métricas, podemos considerar el grado de acuerdo como *sustancial* ($0.61 < \kappa < 0.80$), lo cual puede

	Evaluadores				
Uso*	**1**	**2**	**3**	**Voto**	**C**
acné	Contr	No-Contr	Contr	**Contr**	-1.441
depresión	Contr	Contr	Contr	**Contr**	2.372
control de natalidad	Contr	Contr	Contr	**Contr**	-2.889
obesidad	Contr	Contr	Contr	**Contr**	3.023
desorden bipolar	No-Contr	No-Contr	No-Contr	No-Contr	-3.221
ansiedad	No-Contr	Contr	No-Contr	No-Contr	3.331
TDAH	No-Contr	No-Contr	No-Contr	No-Contr	4.890
estreñimiento	Contr	Contr	Contr	**Contr**	-5.636
dejar de fumar	No-Contr	No-Contr	No-Contr	No-Contr	-7.084
candidosis vaginal	No-Contr	No-Contr	No-Contr	No-Contr	-22.026

Tabla 4.9: Resultados del ciclo de anotación realizado por tres anotadores independientes. Se obtiene un grado de acuerdo del .866 y k =.733. La columna "voto" representa el voto mayoritario y adicionalmente presentamos la columna "C", ordenada de mayor a menor controversia (de menor a mayor valor absoluto, dada por la definición). (*La columna "Uso" representa la combinación fármaco+uso; se utiliza esta última por comodidad.)

acuerdo	κ
.866	.733

Tabla 4.10: Grado de acuerdo entre anotadores, para las anotaciones presentadas en la tabla 4.9

manifestar una buena definición de la tarea de anotación, que ha sido desarrollada de manera muy similar por todos los anotadores. Igualmente, podemos considerar que el voto mayoritario de los anotadores, en este grado de acuerdo, es una buena medida con la que comparar los resultados obtenidos mediante nuestro *pipeline*.

4.2.4. Análisis Cualitativo y Discusión de los Resultados.

En primer lugar, comparando en la tabla 4.9 los resultados del voto mayoritario de los anotadores con los resultados obtenidos por el sistema (columna **C**), podemos apreciar que si ordenamos de acuerdo al valor de la estimación de controversia[6] obtenido mediante nuestro *pipeline*, aquellos casos que se encuentran en la primera mitad de la tabla coinciden con los casos para los que los evaluadores decidieron que existía controversia, salvo en un caso. Asimismo, aquellos que han sido clasificados como "no-controvertidos" han sido estimados mayoritariamente como poco controvertidos por el sistema.

Analizaremos a continuación, de manera cualitativa, los diferentes escenarios de acuerdo, desacuerdo que se han dado entre sistema y anotadores, así como los desacuerdos internos entre anotadores. Para dicho análisis, se han extendido los

[6] en el cual, un mayor | C | representa menor controversia y C= 0 máxima controversia.

resultados de la tabla 4.9 en la tabla 4.11, incluyendo el detalle del acuerdo absoluto entre anotadores columna **AA** (Acuerdo Absoluto), como otra dimensión interesante que analizar en nuestros resultados.

Para llevar a cabo el análisis cualitativo de los resultados, se han utilizado las impresiones, *a posteriori*, que la tarea de anotación ha causado en los anotadores, así como la definición de controversia (sección 3.1.1) y diferentes representaciones gráficas de los *clusters* obtenidos en el proceso de agrupación, junto con sus polaridades.

Un aspecto adicional a tener en cuenta para el análisis, extraído a partir del *feedback* de los anotadores, es que los *clusters*, o sub-temas que se generan en cada caso de uso, tienden a representar aspectos del uso del fármaco como efectos secundarios o grupos de efectos secundarios.

	Evaluadores					
Uso*	**1**	**2**	**3**	**Voto**	**AA**	**C**
acné	Contr	No-Contr	Contr	**Contr**	✗	-1.441
depresión	Contr	Contr	Contr	**Contr**	✓	2.372
control de natalidad	Contr	Contr	Contr	**Contr**	✓	-2.889
obesidad	Contr	Contr	Contr	**Contr**	✓	3.023
desorden bipolar	No-Contr	No-Contr	No-Contr	No-Contr	✓	-3.221
ansiedad	No-Contr	Contr	No-Contr	No-Contr	✗	3.331
TDAH	No-Contr	No-Contr	No-Contr	No-Contr	✓	4.890
estreñimiento	Contr	Contr	Contr	**Contr**	✓	-5.636
dejar de fumar	No-Contr	No-Contr	No-Contr	No-Contr	✓	-7.084
candidosis vaginal	No-Contr	No-Contr	No-Contr	No-Contr	✓	-22.026

Tabla 4.11: Resultados de evaluación ampliados. Se obtiene un grado de acuerdo del .866 y k =.733. La columna "voto" representa el voto mayoritario y adicionalmente presentamos la columna "C", ordenada de mayor a menor controversia. La columna "AA" representa el acuerdo absoluto de los anotadores (en rojo, encontramos los casos con desacuerdos entre anotadores). *La columna "Uso" representa la combinación fármaco+uso; se utiliza esta última por comodidad.

4.2.4.1. Desacuerdo entre anotadores

Analizando los casos en los que los anotadores no han estado en completo acuerdo (tabla 4.11), podemos detectar dos casos principales en los que los anotadores han aplicado de manera diferente el criterio de anotación. Consideraremos estos casos de interés para su análisis, como aquellos que provocan mayores dudas en la anotación.

El caso del fármaco para **acné** ha sido anotado mayoritariamente como controvertido, con el desacuerdo de uno de los anotadores. En el conjunto de argumentos, encontramos un escenario de gran polarización, donde las informaciones que se proporcionan en los diferentes argumentos parecen contradecirse y no podemos discernir si aspectos concretos del fármaco provocan efectos positivos o negativos.

En este caso, el fármaco provoca una serie de efectos secundarios comunes para la mayoría de los usuarios, variando en gravedad, pero si se sigue con el uso del fármaco durante el tiempo necesario recomendado, los efectos secundarios remiten y se obtienen buenos resultados. Esto resulta en una situación en la que los usuarios que no han persistido en el uso del fármaco y lo abandonan antes de tiempo, se quedan solamente con los efectos negativos del fármaco y nunca llegan a ver resultados.

Al tratarse de una situación de consenso, en la que la mayoría de los usuarios describen un escenario común, solamente en dos fases diferentes del tratamiento, uno de los anotadores ha extrapolado falta de controversia al deducir estos efectos negativos como una primera fase de un proceso que aplica a todos los usuarios.

Los demás anotadores indican, que han observado un comportamiento similar, pero no han usado dicho criterio tras haberse concretado la matización del criterio de anotación en el paso *(4)* del proceso de anotación (sección 4.2.3).

Bajo el criterio de anotación matizado, este caso contiene claramente indicios de controversia. En primer lugar, no se evidencia una ausencia de controversia. En segundo lugar, considerando exclusivamente la información disponible en dichos comentarios, sin aplicar dicho proceso deductivo, la situación consiste en grupos de usuarios que describen experiencias opuestas respecto al uso del fármaco.

El criterio usado por el anotador en desacuerdo, se basa en un proceso deductivo, pero no en la información disponible en el texto de manera explícita. Para llegar a tales conclusiones, deberían desarrollarse, partiendo de este marco, diversas técnicas de *minería de controversia*, que nos permita extraer las razones por las que un caso es controvertido. Un desarollo de este marco en esta dirección, sería una línea potencial de trabajo futuro.

Respecto al *pipeline*, tal grado de controversia obtenido por la estimación del sistema puede venir dado por la presencia de un gran cluster fuertemente polarizado negativamente junto a una serie de *clusters*, de tamaños relevantes en comparación, polarizados positivamente que constituyen una buena representación de confrontación de diferentes aspectos positivos frente a un gran aspecto negativo (ver Figura 4.2). Dicho cluster con carga predominantemente negativa, representa el conjunto de argumentos de aquellos usuarios que sólamente han comprobado los efectos secundarios del principio del tratamiento y por tanto, sólo pueden expresarse negativamente hacia el uso del fármaco.

El otro caso que encontramos en el que uno de los anotadores ha discrepado de los otros dos, es el caso del fármaco para la **ansiedad**. En este caso, el anotador identifica una población mayoritaria para la que el fármaco funciona correctamente, frente a una población minoritaria con efectos secundarios muy graves. En este caso, la gravedad de los síntomas ha influenciado en el criterio del anotador, considerando estos dos aspectos igualmente relevantes y por tanto, dándose cierta controversia. Para los otros dos anotadores, sin embargo, la opinión sobre el fármaco está clara-

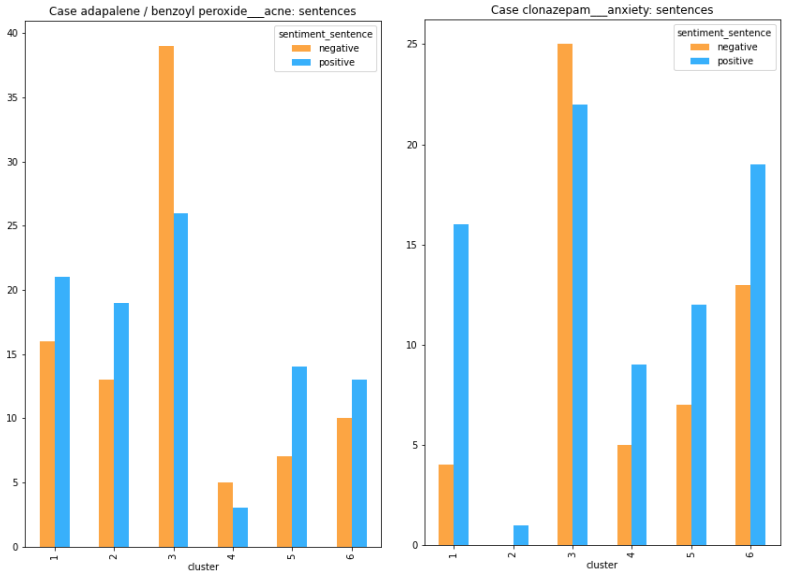

Figura 4.2: Casos considerados para los que los evaluadores no han alcanzado un acuerdo absoluto. Izquierda: caso *acné*; derecha: caso *ansiedad*. El eje horizontal representa los diferentes *clusters* y el eje vertical número de argumentos. La clasificación de sentimiento se muestra en azul y naranja, para positivo y negativo, respectivamente.

mente consensuada, encontrando los diferentes aspectos del fármaco con polaridad claramente positiva.

Para el sistema, este caso guarda un cierto patrón con el caso anterior del fármaco contra el acné. En ambos casos se constituye el mismo número de *clusters* con contenido semántico similar, con una distribución de tamaños y polaridades similar. En ambos casos encontramos un cluster de gran tamaño con polaridad neta negativa y un grupo de *clusters* de tamaños relevantes, con polaridades claramente positivas. La diferencia entre los dos radica en el equilibrio de la polaridad del cluster mayor frente al conjunto de polaridades positivas de los *clusters* menores. En el caso del acné este equilibrio se mantiene, resultando en un caso controvertido, mientras que en el caso del fármaco para ansiedad, este equilibrio se rompe ampliamente hacia la polaridad positiva, obteniendo una estimación que podríamos considerar como *no-controvertida*, que por otra parte ha sido el voto mayoritario de los anotadores.

Sin embargo, ante la falta de un *threshold*, este caso deberá considerarse como un caso umbral de nuestro sistema, en el que no podemos garantizar la pertenencia a uno u otro lado de la clasificación.

Estos casos serán analizados en los siguientes apartados.

4.2.4.2. Casos claros de Controversia para los Evaluadores y para el Sistema

Se consideran casos claros de controversia tanto para el sistema como para los evaluadores, los casos de **depresión**, **obesidad** y **control de natalidad**, que además han gozado de un acuerdo interno absoluto.

Tanto en el caso del fármaco para la **depresión** como en el caso del fármaco para la **obesidad**, los evaluadores perciben una población relevante a la que el fármaco le funciona, otra población a la que no le funciona, y una serie de efectos secundarios variados y en ocasiones, contradictorios.

En ambos casos, son igualmente identificados como casos controvertidos por nuestro sistema, que detecta la existencia de muchas sub-temáticas semánticamente diferentes, de tamaños muy similares, sin que se imponga de manera relevante ninguna de las dos polaridades (ver Figura 4.3). Esta distribución de los sub-temas de manera homogénea en tamaño y polaridad, es interpretada por el sistema como controvertida, ante una falta de consenso hacia una de las posiciones. Puede observarse que en el caso del fármaco para la obesidad, existen inclinaciones más relevantes hacia el aspecto positivo de las temáticas (*clusters* 1 y 5 en la figura 4.3, izquierda). Esto se traduce en el cómputo como un grado de controversia menor, al haber aspectos del fármaco que son considerados claramente positivos.

En el caso de **control de natalidad**, se observa un número menor de *clusters*, de tamaños mayores que los presentados en los casos anteriores. En este caso, las sub-temáticas, según han confirmado los anotadores, se correspondían con sub-casos de uso del fármaco y no directamente con grupos de efectos secundarios. Esto indica, que en este caso tenemos un nivel más de jerarquía (caso, sub-caso, efectos secundarios), con respecto a los otros casos, que suelen presentar una jerarquía más simple (caso, efectos secundarios). Esto se debe a los múltiples usos (regulación hormonal, problemas de piel, problemas de peso, etc.) de la píldora anticonceptiva en la comunidad de usuarios, aparte de su propósito original.

Nuestro sistema, por tanto, ha estimado la controversia basándose en la distribución, tamaño y polaridad de los sub-casos de uso, al constituir grupos semánticamente coherentes. Dicha estimación da como resultado un grado de controversia claro, de polaridad general negativa (C =-2.889). Puede verse en la figura 4.3, como las diferencias de polaridades positivas y negativas son relevantes, imponiéndose ligeramente la negativa. Esto podría ser interpretado como cuatro sub-casos de uso del fármaco, para los que, en tres de ellos, se imponen ligeramente los aspectos negativos, frente a un caso de uso claramente positivo, pero con menor presencia.

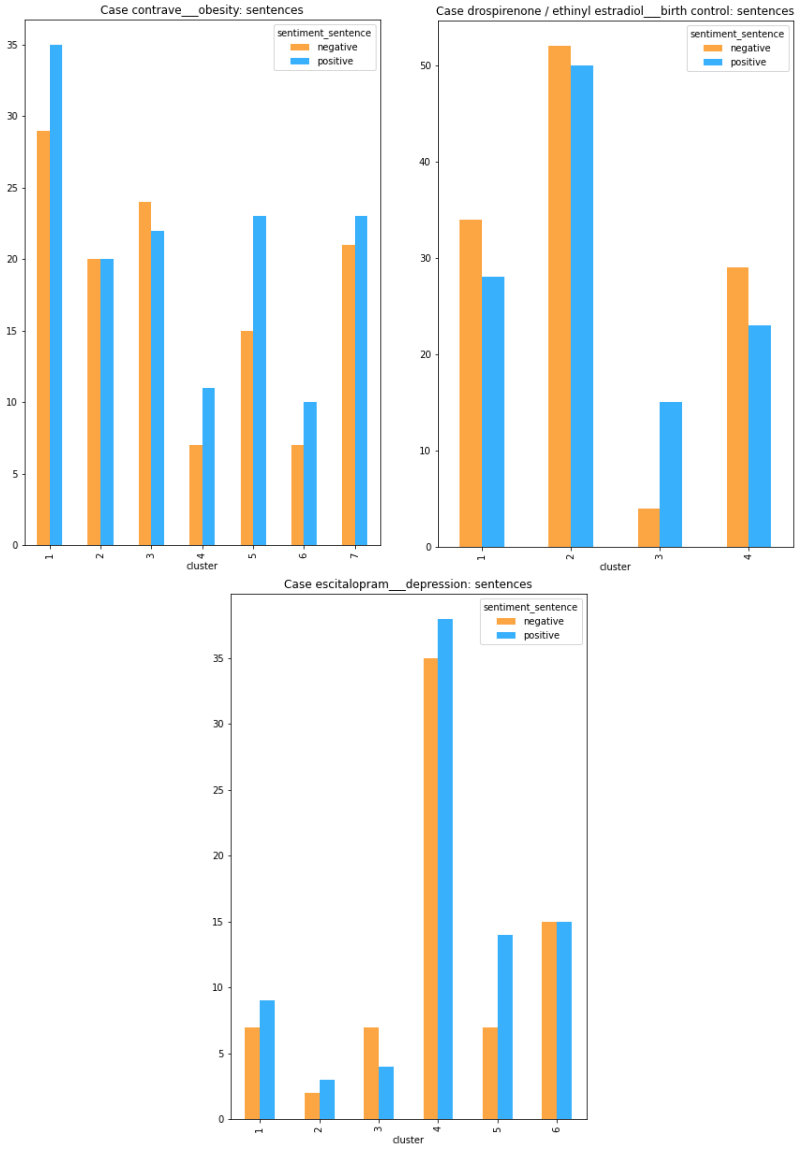

Figura 4.3: Casos considerados controvertidos por el sistema y por los evaluadores. Izquierda: caso *obesidad*; derecha: caso *control de natalidad*. El eje horizontal representa los diferentes *clusters* y el eje vertical número de argumentos. La clasificación de sentimiento se muestra en azul y naranja, para positivo y negativo, respectivamente.

4.2.4.3. Casos claros de No-Controversia para los Evaluadores y para el Sistema

Existen tres casos para los que claramente los evaluadores han decidido que no se da controversia y esto ha sido confirmado por la estimación proporcionada por el

sistema, siendo los últimos de la lista.

El caso del fármaco para **TDAH**, ha resultado bastante claro para los evaluadores, que han comprobado que en cada uno de los aspectos que se dan en los argumentos, la posición positiva se impone a la negativa, que suele ser expresada en forma de efectos secundarios desagradables pero soportables y comunes (falta de apetito, pérdida de peso, síndrome de abstinencia, etc.), a cambio de unos buenos resultados y beneficios en general.

Esto puede comprobarse en los resultados del sistema (ver Figura 4.4), el cual captura la existencia de tres sub-temáticas principales de argumentos, en los que el aspecto positivo siempre se impone al negativo. El cómputo de la estimación de controversia en estas condiciones resulta en un valor relativamente elevado y positivo de C.

En el caso del fármaco para **dejar de fumar**, para los evaluadores está, de nuevo, claro el consenso de una mayoría de usuarios que coinciden en los resultados positivos del fármaco y cómo los efectos secundarios son aceptables.

Sin embargo, nuestro sistema, a pesar de coincidir en la una estimación de baja controversia, lo ha obtenido por razones equivocadas. En este caso, se da particularmente una gran cantidad de comentarios en los que se aportan experiencias negativas (comparación con otros productos, situación personal, intentos de dejar de fumar en el pasado, etc.) en contraste con los buenos resultados gracias al uso de este fármaco. En esta situación, la estimación de controversia ha funcionado correctamente, pero sobre el conjunto de elementos equivocado, ya que estas experiencias argumentativas negativas han sido consideradas en el cómputo de la controversia. Debe profundizarse en el filtrado y minería de los elementos argumentativos presentes, si se quieren excluir estos elementos de la estimación de controversia.

Adicionalmente, hemos podido detectar una anomalía en los resultados, gracias al signo de C, que representa la polaridad global del caso. Los evaluadores han reportado un consenso en sentido positivo hacia el fármaco mientras que el valor de la controversia indicaba baja controversia (alto consenso), pero de polaridad negativa ($C = -7.084$). Esta polaridad no se ha tenido en cuenta de manera formal en los ciclos de evaluación y es una posible mejora futura.

En el caso del fármaco para **candidosis vaginal**, hay un clarírisimo consenso general sobre los efectos negativos de este fármaco, identificado tanto por los evaluadores como por el sistema. Si revisamos los resultados del sistema en la figura 4.4, existen diversos *clusters* que manifiestan aspectos con una diferencia de polaridad muy reducida, y por otro lado, un gran *cluster* con polaridad negativa, que hace que se dispare el valor de C hasta -22.026, indicando una situación muy alejada de la controversia.

Podemos entender mejor las capacidades de nuestro sistema si comparamos este caso junto a otro caso que puede resultar similar, *a priori*, pero que se trata en realidad de un caso controvertido (ver Figura 4.5).

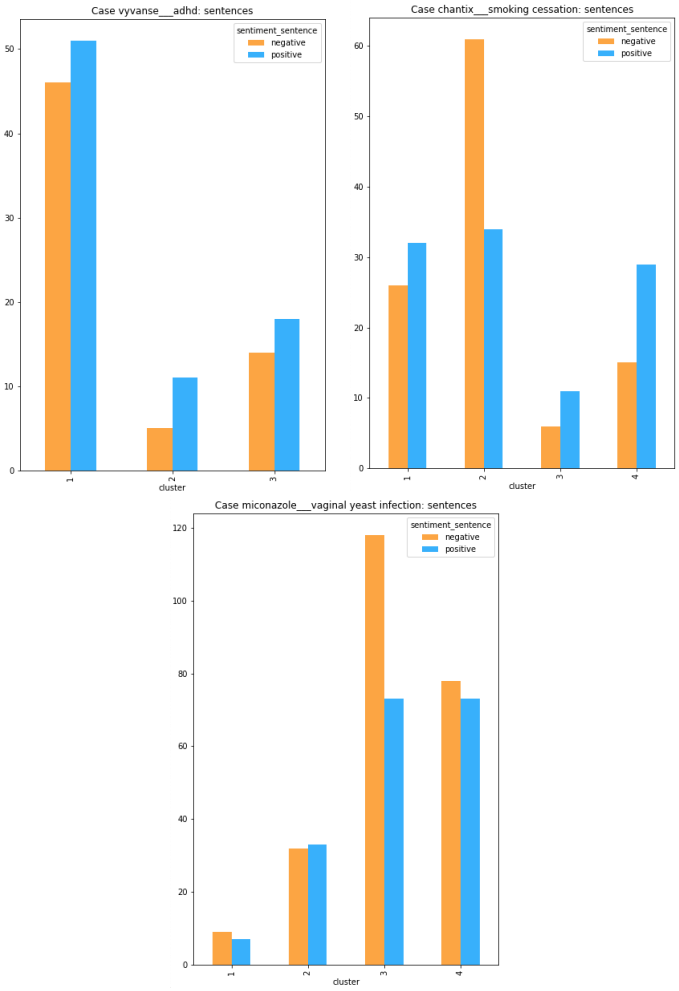

Figura 4.4: Casos considerados claramente no-controvertidos por el sistema y por los evaluadores. Arriba izquierda: caso *TDAH* (AHDH en inglés); arriba derecha: caso *dejar de fumar*; abajo: caso *candidosis vaginal*. El eje horizontal representa los diferentes *clusters* y el eje vertical número de argumentos. En azul aquellos clasificados como positivos y en naranja, aquellos considerados como negativos.

Podemos observar una distribución de *clusters* muy similar entre ambos casos, encontramos el mismo número de *clusters*, con una distribución de tamaños muy parecida, representando las diferentes sub-temáticas de cada caso. En el caso de control de natalidad los *clusters* con polaridad neta negativa son compensado en gran medida por el *cluster* en el que se impone una clara polaridad negativa neta (el número 3 en la figura 4.5, derecha). El resultado de este escenario, según nuestra defínción de C (sección 3.2.5), es un valor de C relativamente bajo y negativo (-2.889), lo cual indica que la situación final es controvertida, con una polaridad global negativa. En un caso práctico de ejemplo, un usuario interesado en este fármaco, podría leer estos comentarios para construirse una opinión propia, antes de decidirse a consumirlo, y no tendría nada clara su decisión, al no percibir una situación de consenso lo suficientemente clara.

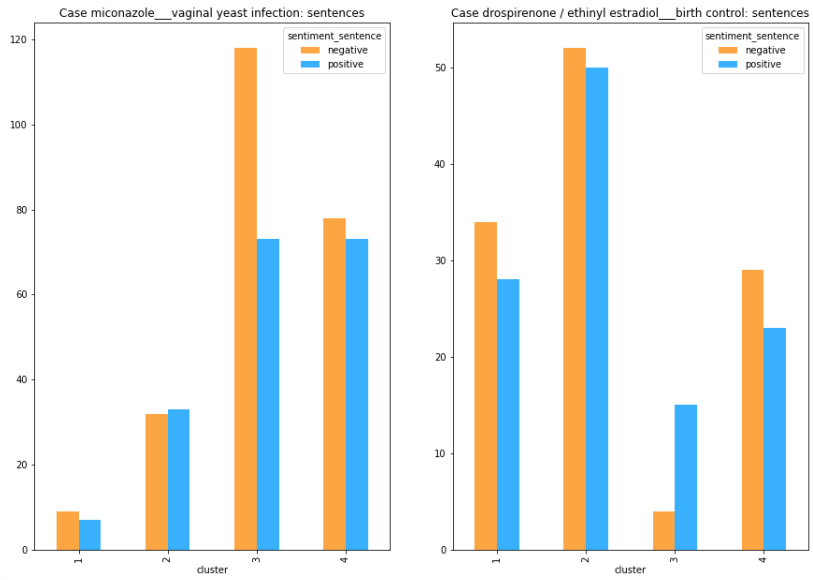

Figura 4.5: Comparación de casos de distribución de *clusters* similares. Izquierda: caso *candidosis vaginal*; derecha: caso *control de natalidad*; abajo: caso *acné*. El eje horizontal representa los diferentes *clusters* y el eje vertical número de argumentos. En azul aquellos clasificados como positivos y en naranja, aquellos considerados como negativos.

Volviendo a nuestra comparación, en el caso del fármaco para candidosis vaginal, observamos que a pesar de que las polaridades de los argumentos están distribuidas de manera equilibrada en cada sub-temática, existe un cluster (el número 3) en el que los argumentos negativos están mucho más presentes que los positivos, siendo además el cluster de más tamaño. Esta masa de argumentos negativos no puede ser compensada por ninguna de las otras temáticas en sentido opuesto, por lo que el resultado final será un valor de C muy alto y negativo (-22.026), repre-

sentando un amplio consenso hacia los aspectos negativos del fármaco. Un usuario que leyese estos argumentos, en este caso, no tendría ninguna duda.

En ambos casos, los resultados predichos por el sistema y el análisis cualitativo realizado mediante la lectura de las figuras ha anticipado una realidad confirmada por las anotaciones y el *feedback* de los anotadores.

4.2.4.4. Casos de No-Controversia para los Evaluadores y umbrales para el Sistema

Consideremos aquellos casos que son considerados como no-controvertidos por los evaluadores, pero al no estar en los extremos de nuestro intervalo de resultados y al no tener un umbral *a priori*, dando el valor relativo de la estimación de controversia, no podemos interpretar si son controvertidos o no, desde el punto de vista del sistema.

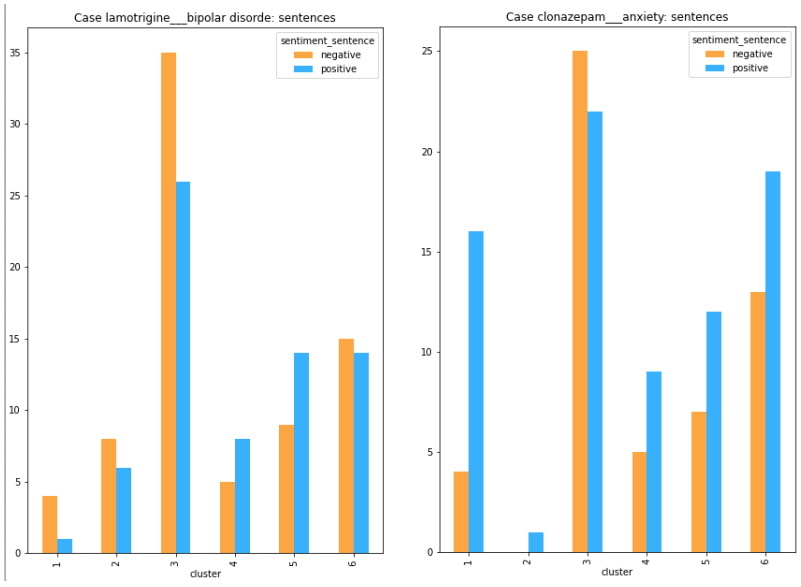

Figura 4.6: Casos considerados como no-controvertidos por los evaluadores pero casos *umbral* por el sistema. Izquierda: caso *desorden bipolar*; derecha: caso *ansiedad*. El eje horizontal representa los diferentes *clusters* y el eje vertical número de argumentos. En azul aquellos clasificados como positivos y en naranja, aquellos considerados como negativos.

Para los evaluadores también existen algunas ligeras dudas, ya que no ha existido acuerdo absoluto, respecto al caso de **ansiedad** por ejemplo, aunque el voto mayoritario ha resultado en *no-controvertido*. La razón radica en que una población mayoritaria considera, en general, positivos los diferentes aspectos del fármaco, pero existe una pequeña población cuyos efectos secundarios son de gravedad. Esto

puede afectar a la evaluación y que el evaluador le dé un mayor peso a este conjunto, generando una situación de controversia leve.

Observando la figura 4.6, observamos un escenario donde el conjunto de *clusters* de polaridad neta positiva es más relevante que aquel en el que exite una inclinación hacia la polaridad negativa. Diríamos en este caso que estamos más cerca de una situación de no-controversia, que de una situación de controversia.

Para el caso del fármaco para **desorden bipolar**, los evaluadores detectan un consenso muy amplio en los diferentes aspectos del fármaco, inclinado hacia el lado positivo, con pequeños grupos de casos raros o graves.

En nuestro sistema, sin embargo, obtenemos una estimación de controversia leve, pero con polaridad negativa ($C = -3.221$). Observando la figura 4.6, podríamos concluir que es un caso potencialmente más controvertido que no-controvertido, inclinado hacia un ligero consenso en los aspectos negativos. Lo que está ocurriendo, de nuevo, al igual que en el caso del fármaco contra el tabaquismo, es que estamos incluyendo en el proceso una masa de argumentos negativos derivados de la descripción de experiencias negativas pasadas de los usuarios, en contraste con unos resultados aceptables del fármaco en cuestión. De nuevo, surge la necesidad de realizar una *minería de la controversia* y de profundizar más en el procesamiento de los elementos argumentativos, para discernir entre experiencias pasadas, actuales, comparaciones, etc.

A raíz de estas observaciones, puede plantearse, además, el debate del grado de controversia de un tema, de si la controversia debe ser una señal no-binaria, y especialmente si debería generarse un *output* de tipo cualitativo, como puede ser un resumen de los argumentos a favor y en contra de un tema. En este sentido, además de aportar una información muy valiosa, tendríamos otra vía para su evaluación.

4.2.4.5. Caso controvertido para los Evaluadores, pero no para el Sistema

En esta categoría, tenemos solamente el caso del fármaco contra el **estreñimiento**, un caso que consensuadamente ha sido considerado controvertido por los evaluadores pero como poco controvertido por el sistema. Se trata del único caso de conflicto entre la opinión de los evaluadores y la estimación proporcionada por el sistema, ya que en el resto de casos se aprecian resultados coherentes entre ambas partes.

Si analizamos en profundidad este caso, encontramos un caso de un tipo de controversia no considerada por nuestro sistema.

Para los evaluadores, se trata de un caso donde el conjunto de efectos secundario coincide parcialmente con el conjunto de efectos que son parte del funcionamiento usual del fármaco. Algunos de los efectos son dolores abdominales, calambres y movimiento intestinal.

La principal causa de la falta de consenso radica en la intensidad de dichos

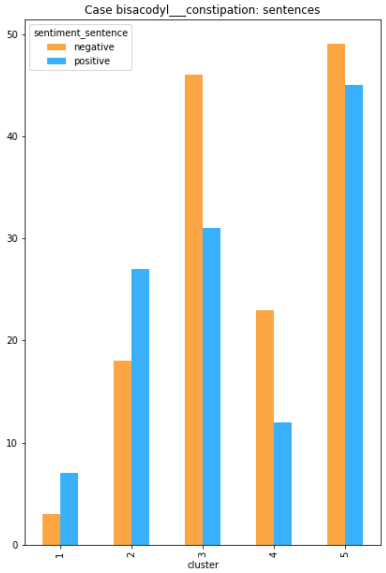

Figura 4.7: Caso controvertido para los evaluadores, pero no-controvertido según el sistema. El eje horizontal representa los diferentes *clusters* y el eje vertical número de argumentos. En azul aquellos clasificados como positivos y en naranja, aquellos considerados como negativos.

efectos. Para algunos, estos efectos son un paso soportable y necesario dentro de un funcionamiento correcto el fármaco. Para otros, estos efectos alcanzan una intensidad insoportable o su resultado final es incontrolable y desastroso, acabando algunos de ellos en el hospital.

Sin embargo, en nuestro sistema (ver Fig.4.7), esto es interpretado como diferentes sub-temáticas en las que mayoritariamente se impone el aspecto negativo y un par de sub-temáticas de menor tamaño en las que el aspecto positivo se impone, pero aún así hay una posición general mayoritariamente negativa hacia el medicamento, cercano al consenso.

Capítulo 5

Conclusiones y trabajo futuro

En este capítulo, hacemos una recopilación de las diferentes conclusiones extraídas durante y al final del desarrollo de este trabajo, así como de los hitos que se han llevado a cabo y las implicaciones de estos. A continuación se exploran y proponen diferentes líneas de trabajo futuro, inspiradas por los resultados y conclusiones del trabajo realizado.

5.1. Trabajo realizado

En este trabajo se ha realizado un estudio del problema de la detección de controversia, identificando cuáles son los desafíos de las metodologías existentes en el estado del arte para este problema. Entre estos desafíos, encontramos una falta de definición explícita y ampliamente aceptada y aplicada, así como una metodología para su detección acordemente amplia e independiente del dominio y caso de uso. Para afrontar dichos desafíos, hemos desarrollado una propuesta para una definición amplia de controversia y una aproximación técnica para su detección, además de su implementación y evaluación en un caso de estudio concreto: el de comentarios de usuarios en foros del ámbito médico (corpus *Drug Review Dataset*).

Tras un extenso análisis del estado del arte, se ha identificado una profunda dependencia del concepto de controversia de los ámbitos, técnicas, contenido y plataformas en los que se han ido desarrollando las diferentes aproximaciones. Esta dependencia dificulta el desarrollo de un marco teórico adecuado alrededor de un concepto *amplio* de controversia, aplicable para los diferentes ámbitos en los que pueda ser de utilidad y susceptible de ser fácilmente adoptado, al incluir los conceptos claves de las propuestas existentes.

Nuestro primer objetivo ha sido construir dicho marco en torno a una defi-

nición amplia de controversia de dichas características, capturando los principales conceptos de las metodologías presentadas en el estado del arte y sin perder el significado convencional de lo que constituye la controversia para un usuario ajeno a este trabajo. Entre estas condiciones se han extraído: *(1)* una estructura argumental implícita; *(2)* la formación de posiciones de individuos o grupos de individuos respecto a un tema; *(3)* la presencia de un sentimiento o polaridad y *(4)* finalmente la contraposición de dichas posiciones.

De estos puntos se ha derivado una definición, sobre la que construir, paso a paso, un sistema de detección de controversia modular que aborde, una por una, las condiciones establecidas por la definición. Según nuestra definición, consideraremos controvertido todo aquel tema, que provoque una confrontación de grupos de opinión con polaridades opuestas, a través de mecanismos argumentativos. Y el grado de controversia asociado a dicho tema será mayor o menor, dependiendo de ciertas características de esta confrontación, como pueden ser el tamaño, distribución y grado de polarización de los grupos de opinión mencionados.

Para afrontar estar tarea hemos propuesto un sistema de detección modular, compuesto por tres componentes: un Clasificador Genérico, un Clasificador de Dominio y una Señal de Dominio (presencia de síntomas, en este caso).

Para el primer clasificador, Clasificador Genérico, hemos comprobado la validez del uso de *features* de un corpus y dominio particular (*Persuasive Essays*)[1] en el caso de un corpus de amplio dominio *Cross-Domain Setential Argument Mining*[2], mejorando los resultados del estado del arte para esta tarea y corpus ($F1$ =.73 frente a $F1$ =.68). El Clasificador Genérico es capaz de identificar aquellos argumentos que utilizan mecanismos generales de argumentación, independientemente del dominio. Sin embargo, no ha resultado suficiente para detectar satisfactoriamente argumentación en este dominio.

Para capturar aquellos argumentos que poseen unas características de argumentación muy ligadas al dominio, se han seleccionado, a través del análisis cualitativo del corpus *Drug Review Dataset*, un conjunto de *features* de dominio, consideradas relevantes para la argumentación (elementos del ámbito médico, posologías, síntomas, etc.) y se han utilizado para construir un Clasificador de Dominio utilizando un subset de *Drug Review Dataset* como set de entrenamiento.

Estos dos clasificadores se han combinado, junto a una Señal de Dominio (presencia de síntomas) que se ha considerado de gran relevancia en este caso, a través de un voto mayoritario para obtener una señal final de clasificación, con unas garantías de precisión (P =.92 para la clase *Argument*) y *performance* ($F1$ =.68, superior a nuestro *baseline* no-informativo), suficientemente sólidas para construir el resto de los pasos del sistema, basados en una buena predicción de la clase *Argument*.

Los argumentos detectados por el componente anterior, son agrupados semánti-

[1] (Stab y Gurevych, 2014a)
[2] (Stab et al., 2018)

camente mediante el uso de *word-embeddings* y un algoritmo de *clustering* aglomerativo jerárquico. Los grupos generados durante este proceso representarán las diferentes sub-temáticas o aspectos presentes en el tema de controversia considerado. A continuación, se ha clasificado la polaridad de los argumentos que conforman cada sub-tema, mediante un modelo de clasificación de polaridad del estado del arte.

Basándonos en la distribución, tamaño y grado de polaridad de dichas subtemáticas, se realiza una estimación del grado de controversia, acorde a la definición propuesta.

Para la evaluación de nuestro caso particular, se ha procesado en nuestro sistema un conjunto de instancias del dataset *Drug Review Dataset*, constituido por comentarios de usuarios sobre fármacos, en el que cada comentario constituye un documento para nuestro sistema, a segmentar y procesar como oraciones. Nuestro objeto de controversia en este caso son los diferentes casos de uso de fármacos a sintomatologías específicas (por ejemplo, *aspirina+migrañas*).

De los resultados, se extraído un conjunto de 10 casos de uso, que han sido manualmente anotados de manera independiente por tres anotadores diferentes, a los que se le ha proporcionado nuestra definición de controversia y los cuales han alcanzado un alto grado de acuerdo en sus anotaciones.

A partir de los resultados y la anotación manual, se ha realizado un análisis cualtitativo y comparativo de los resultados.

5.2. Conclusiones

Para finalizar, hemos de destacar que se han cumplido con los objetivos que nos habíamos marcado. En primer lugar, hemos desarrollado una **definición de controversia** que extrae y aglutina los conceptos principales de la controversia, presentes en el estado del arte, de manera constructiva, compatible, estructurada y fácilmente aplicable como metodología. Esta definición supera uno de los desafíos del problema de la detección de controversia, el de desarrollar un marco que sea independiente del dominio y de la forma y origen de los documentos, frente aquellas aproximaciones del estado del arte dependientes de las características de plataforma (*Twitter*, *Wikipedia*, etc.), de dominio (ensayos, *tweets*, artículos, etc.) o de un aspecto concreto de la controversia (sentimiento, discusión, estructura de grafo, etc.)

Nuestra aproximación es además novedosa en la aplicación formal de la **detección de argumentos**, como base para detectar y estimar la controversia de un tema.

Para este componente en particular, hemos conseguido resultados de detección genérica de argumentación que superan los del estado del arte en esa tarea y corpus ($F1$ =.73 frente a $F1$ =.68).

Se ha demostrado que la implementación de la definición en forma de un **sistema de detección de controversia** es viable y se ha demostrado su aplicación con éxito en el caso particular del corpus *Drug Review Dataset*. Durante esta aplicación, se ha comprobado que su modularidad es una de sus grandes ventajas, pudiendo incluir *inputs* y *features* relevantes, propios del dominio particular, de manera simple y efectiva.

Los resultados obtenidos son prometedores, ya que el sistema es capaz de detectar satisfactoriamente la controversia en sus extremos. Es decir, es capaz de reconocer aquellos casos que son muy poco controvertidos, de aquellos que son muy controvertidos. Por otro lado, los casos intermedios han resultado difíciles de clasificar, tanto para el sistema como para los anotadores, ya que la controversia no es de carácter binario y su término medio está lleno de matices.

Como parte de dicha evaluación, hemos sentado las bases de un **marco de evaluación y anotación** para el problema de la detección de controversia, no presente en el estado del arte, y que puede servir de base para trabajos futuros.

Adicionalmente, se ha comprobado mediante un análisis cualitativo y comparativo, que nuestro sistema captura correctamente muchas de las características propias de los diferentes casos de controversia considerados en forma de sub-temas, tamaño de dichos sub-temas, distribución de argumentos, polaridad, etc. Estas observaciones han sido, en su mayoría, contrastadas con el *feedback* cualitativo aportado por los anotadores tras realizar su tarea de evaluación.

Por último, destacamos las capacidades, no sólo de estimación y detección de controversia, sino también de **explicabilidad**, de los resultados de nuestro sistema. Esto puede sentar las bases para trabajos posteriores, en el sentido de desarrollar una *minería de la controversia*, que nos permita obtener resultados cualitativos explicables, que en el caso de la controversia, aportan una información mucho más valiosa y útil que una simple estimación cuantitativa.

5.3. Trabajo futuro

La definición del problema de detección de controversia estudiado en este trabajo, el estudio de las diferentes posibilidades para su resolución, el análisis del estado del arte y finalmente nuestra propuesta de definición, implementación y aplicación de un aproximación como la aquí presentada, han expuesto e inspirado un amplio abanico de ideas y líneas de desarrollo a las que este trabajo podría conducir y facilitar. Algunas de estas líneas de trabajo futuro podrían ser de inmediato desarrollo y ejecución, partiendo de lo presentado en este trabajo y algunas de ellas presentan ideas de innovación, enfocadas más bien en el largo plazo, pero igualmente dignas de mención. Exploramos algunas de ellas:

Posterior desarrollo y mejora de los componentes individuales

Durante el trabajo se ha mencionado en varias ocasiones el hecho de alcanzar unos niveles mínimos aceptables en el *performance* de cada componente, en pos de demostrar la viabilidad completa del sistema propuesto, sin llegar a optimizar cada uno de los componentes hasta el máximo de sus capacidades.

Un primer paso para continuar el trabajo presentado aquí, sería mejorar los resultados obtenidos en cada uno de los componentes, maximizando al mismo tiempo sus capacidades de generalización a otros dominios y casos de uso. Por motivos de limitación de recursos en el marco de este desarrollo, no han podido realizarse aplicaciones del sistema a otros casos de uso o ámbitos para comprobar su capacidad de generalización, como se ha realizado con el caso de *Drug Review Dataset*. Es el caso del componente de detección de argumentación, que ha ganado gran complejidad durante su desarrollo y para el que podrían desarrollarse otras soluciones posibles, incluyendo además otros corpora, *features* y técnicas de aprendizaje máquina.

Desarrollo de corpora para la detección de controversia

Uno de los principales desafíos en la resolución del problema de la detección de controversia es la falta de corpora anotados para este problema. Esta sería, sin duda, una de las líneas de trabajo más factibles y útiles para el futuro, ya que un corpus de estas características sería útil para toda la comunidad, en este y otros problemas.

Hemos comprobado en este trabajo, la dificultad de los anotadores para distinguir casos de controversia leve de casos sin controversia y de la sutilidad de esta controversia en los diferentes ámbitos. En el caso que hemos tratado aquí, han existido elementos no previstos que no se encuentran *a priori* en nuestra idea de controversia, que han resultado de relevancia para los anotadores, como puede ser la intensidad de un síntoma.

Esto demuestra, que, a pesar de que hayamos propuesto y demostrado la viabilidad de un enfoque generalista, es necesario elaborar una guía de anotación con pautas para los diferentes dominios, con sus idiosincrasias particulares.

Una vez existan *corpora* de este tipo, podrían desarrollarse y extenderse metodologías como la presentada en este trabajo con más facilidad y garantías, dado que tendríamos un estándar consistente para su evaluación.

Extensión de la definición de controversia

Se ha propuesto en este trabajo una definición técnica para la estimación de controversia que hemos basado de manera intuitiva en la definición teórica, para generar una cuantificación de manera orientativa, según aquellos aspectos que hemos

considerado relevantes sobre la controversia (tamaño de los clusters, distribución, polaridad, etc.)

Hemos comprobado que dicha definición aporta, en términos generales, unos resultados cuantitativamente orientativos y cualitativamente muy prometedores. Pero esta definición podría extenderse, mejorarse y compararse con otras alternativas para elaborar una métrica estandarizada para el estudio de este problema.

Existen muchas otras variables que podrían tenerse en cuenta a la hora de elaborar una estimación de este tipo, como podrían ser el grado de similitud de los *clusters*, una mayor granularidad de la polaridad, estadísticas sobre el conjunto de *clusters*, cronología de los argumentos, etc.

Inclusión de elementos de la teoría de grafos

En el estado del arte hemos analizado algunas propuestas que basaban su enfoque en la teoría de grafos, muy útil sobre todo cuando el escenario de controversia es una plataforma que provee datos de interacción entre usuarios y cuando existe una estructura de conversación en el caso de uso. Una de las desventajas de esta aproximación es la pérdida de generalidad, ya que no todos los escenarios contemplarán un nivel de interacción tan rico como es el caso de una red social, ni será fácil de obtener el grafo si la plataforma no lo genera de manera automática.

Sin embargo, hay elementos de la teoría de grafos que podrían utilizarse para ampliar la propuesta presentada en este trabajo.

Podemos encontrar relaciones entre sub-temas, relaciones entre términos, identificar cuántos usuarios son parte de la controversia, si existen "fuentes" y "sumideros" de controversia o incluso cómo evoluciona un elemento de información a través de una red de usuarios en términos de controversia.

Desarrollo de una *Minería de la Controversia*

Siguiendo en la línea de extraer información a partir de escenarios controvertidos, podemos considerar que podría desarrollarse un subcampo de *Minería de la Controversia*. Como se ha comentado en la Introducción (capítulo 1) y en el Estado del Arte (capítulo 2), el hecho de que un tema genere controversia puede indicar anomalías en el tratamiento de la información, sesgos, hechos de carácter ambigüo, etc. alrededor de dicho tema.

Identificar el porqué de dicha controversia, cuáles son los sub-temas más frecuentes, qué tipo de usuarios están presentes en ella, qué realidad objetiva se encuentra en la base de dicha controversia, etc. puede resultar factible, partiendo de y ampliando un sistema como el presentado. Mediante el desasrrollo de estas técnicas, podríamos "desbloquear" nuevas fuentes de información, no usadas hasta ahora y mejorar aspectos como la calidad, objetividad y las garantías del contenido en la

Web.

Herramienta contra la desinformación

Como parte del desarrollo de una *Minería de la Controversia*, encontramos una aplicación inmediata, y es el uso de su detección y cuantificación para mejorar la calidad y objetividad del contenido en la Web.

Existen hoy en día escenarios, como son las redes sociales, en los que los usuarios consumen y comparten información sin ninguna verificación de su veracidad o calidad, ya sea visible o invisible para el usuario.

Si desarrollásemos un sistema funcional de detección y cuantificación de controversia, dicha cuantificación podría usarse para desarrollar filtros de contenido, mejorar *rankings* e incluso indicar explícitamente al usuario de que la información que va a consumir es de carácter controvertido, sin un fundamento *factual* claro, entre otras aplicaciones.

Una mejora en los Motores de Búsqueda Web

Como una extensión de la aplicación anterior, podríamos incluir estimaciones de controversia en el funcionamiento habitual de los Motores de Búsqueda, a la hora de rastrear e indexar información de la Web, e igualmente en las fases de *ranking* y presentación de los resultados.

Actualmente, un concepto tal como la controversia no es tenido en cuenta a la hora de ofrecer resultados de búsqueda. Esto conduce a varios escenarios en los que podemos recibir información de manera sesgada.

Por ejemplo, temas que no son muy controvertidos, pero su único aspecto de controversia está muy presente en la Web; temas que son muy controvertidos, pero cuya información es de carácter oficialista que minimiza la presencia de dicha controversia en la Web o la dependencia de la controversia de un tema respecto de la región geográfica, entre otros.

Incluyendo este tipo de mecanismos podría ofrecerse un *ranking* más "equilibrado" de resultados, que ayude a minimizar la creación de burbujas de información en el uso de los buscadores web.

Herramienta de análisis de *feedback* de usuarios

Como consecuencia del desarrollo de una *Minería de la Controversia* y en el marco de un ámbito como el estudiado en este trabajo (el de fármacos), puede vislumbrarse una potencial aplicación al problema de extraer *feedback* de manera eficiente y representativa acerca del uso de un producto y acelerar la detección de anomalías en el uso de dicho producto. Esto sería de gran utilidad en el sector médico, para

detectar patrones no tan visibles, de casuísticas en las que una medicación no está funcionando como debía. Igualmente, puede extrapolarse a otros ámbitos y usos, como el de desarrollar nuevas técnicas de *Marketing*, basado en unos *insights* de estas características sobre una comunidad de usuarios.

Herramienta para estudios sociológicos

Considerando un caso más general y menos dirigido que la aplicación anterior, podemos considerar la detección, cuantificación y *Minería de la Controversia* como una herramienta de estudio sociológico más, como hoy en día se utilizan otras como la polaridad que un tema causa en las redes. En ese sentido, el concepto de controversia aportaría mayores detalles sobre ese tipo de escenarios, ya que nuestro concepto engloba al mismo tiempo polaridad, grupos temáticos o de opinión y el hecho de que tenga que existir una estructura argumentativa detrás de lo que se dice.

Bibliografía

Bibliografía

[Addawood y Bashir2016] Addawood, Aseel y Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. En *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, páginas 1–11, Berlin, Germany, Agosto. Association for Computational Linguistics.

[Aharoni et al.2014] Aharoni, Ehud, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, y Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. En *Proceedings of the first workshop on argumentation mining*, páginas 64–68.

[Aker et al.2017] Aker, Ahmet, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, y Mina Ghobadi. 2017. What works and what does not: Classifier and feature analysis for argument mining. En *Proceedings of the 4th Workshop on Argument Mining*, páginas 91–96.

[Andreevskaia, Bergler, y Urseanu2007] Andreevskaia, Alina, Sabine Bergler, y Monica Urseanu. 2007. All blogs are not made equal: Exploring genre differences in sentiment tagging of blogs. En *ICWSM*.

[Azzopardi, De Rijke, y Balog2007] Azzopardi, Leif, Maarten De Rijke, y Krisztian Balog. 2007. Building simulated queries for known-item topics: an analysis using six european languages. En *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 455–462.

[Bach y Moulines2011] Bach, Francis y Eric Moulines. 2011. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. En *Neural Information Processing Systems (NIPS)*, páginas–, Spain.

[Balabantaray, Sarma, y Jha2015] Balabantaray, Rakesh Chandra, Chandrali Sarma, y Monica Jha. 2015. Document clustering using k-means and k-medoids. *arXiv preprint arXiv:1502.07938*.

[Baroni, Dinu, y Kruszewski2014] Baroni, Marco, Georgiana Dinu, y Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. En *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 238–247.

[Bodenreider2004] Bodenreider, Olivier. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

[Bykau et al.2015] Bykau, Siarhei, Flip Korn, Divesh Srivastava, y Yannis Velegrakis. 2015. Fine-grained controversy detection in wikipedia. En *2015 IEEE 31st International Conference on Data Engineering*, páginas 1573–1584. IEEE.

[Carletta1996] Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *arXiv preprint cmp-lg/9602004*.

[Carrillo-de Albornoz et al.2019] Carrillo-de Albornoz, Jorge, Ahmet Aker, Emina Kurtic, y Laura Plaza. 2019. Beyond opinion classification: Extracting facts, opinions and experiences from health forums. *PloS one*, 14(1):e0209961.

[Choi, Jung, y Myaeng2010] Choi, Yoonjung, Yuchul Jung, y Sung-Hyon Myaeng. 2010. Identifying controversial issues and their sub-topics in news articles. En *Pacific-Asia Workshop on Intelligence and Security Informatics*, páginas 140–153. Springer.

[Day y Edelsbrunner1984] Day, William HE y Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24.

[Dori-Hacohen y Allan2013] Dori-Hacohen, Shiri y James Allan. 2013. Detecting controversy on the web. En *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, páginas 1845–1848. ACM.

[Dori-Hacohen y Allan2015] Dori-Hacohen, Shiri y James Allan. 2015. Automated controversy detection on the web. En *European Conference on Information Retrieval*, páginas 423–434. Springer.

[Dori-Hacohen, Yom-Tov, y Allan2015] Dori-Hacohen, Shiri, Elad Yom-Tov, y James Allan. 2015. Navigating controversy as a complex search task. En *SCST@ ECIR*. Citeseer.

[Durant y Smith2006] Durant, Kathleen T y Michael D Smith. 2006. Mining sentiment classification from political web logs. En *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, PA*.

[Florou et al.2013] Florou, Eirini, Stasinos Konstantopoulos, Antonis Koukourikos, y Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. En *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, páginas 49–54.

[Garimella et al.2018] Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, y Michael Mathioukis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.

[Gräßer et al.2018] Gräßer, Felix, Surya Kallumadi, Hagen Malberg, y Sebastian Zaunseder. 2018. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. En *Proceedings of the 2018 International Conference on Digital Health*, páginas 121–125.

[Hatzivassiloglou y Wiebe2000] Hatzivassiloglou, Vasileios y Janyce Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. En *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

[Java y others2007] Java, Akshay y others. 2007. A framework for modeling influence, opinions and structure in social media. En *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volumen 22, página 1933. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[Kamps et al.2004] Kamps, Jaap, Maarten Marx, Robert J Mokken, Maarten De Rijke, y others. 2004. Using wordnet to measure semantic orientations of adjectives. En *LREC*, volumen 4, páginas 1115–1118. Citeseer.

[Kaufman y Rousseeuw1987] Kaufman, Leonard y Peter J Rousseeuw. 1987. Clustering by means of medoids. statistical data analysis based on the l1 norm. *Y. Dodge, Ed*, páginas 405–416.

[Kennedy y Inkpen2006] Kennedy, Alistair y Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125.

[Kim2014] Kim, Yoon. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[Landis y Koch1977] Landis, J Richard y Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, páginas 159–174.

[Lee2019] Lee, Terry. 2019. The global rise of "fake news" and the threat to democratic elections in the usa. *Public Administration and Policy*.

[Linmans, van de Velde, y Kanoulas2018] Linmans, Jasper, Bob van de Velde, y Evangelos Kanoulas. 2018. Improved and robust controversy detection in general web pages using semantic approaches under large scale conditions. En *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, páginas 1647–1650. ACM.

[MacQueen1967] MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. En *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, páginas 281–297, Berkeley, Calif. University of California Press.

[Mikolov et al.2013] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, y Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. En *Advances in neural information processing systems*, páginas 3111–3119.

[Moens et al.2007] Moens, Marie-Francine, Erik Boiy, Raquel Mochales Palau, y Chris Reed. 2007. Automatic detection of arguments in legal texts. En *Proceedings of the 11th international conference on Artificial intelligence and law*, páginas 225–230.

[Palau y Moens2009] Palau, Raquel Mochales y Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. En *Proceedings of the 12th international conference on artificial intelligence and law*, páginas 98–107.

[Pariser2011] Pariser, Eli. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[Popescu y Pennacchiotti2010] Popescu, Ana-Maria y Marco Pennacchiotti. 2010. Detecting controversial events from twitter. En *Proceedings of the 19th ACM international conference on Information and knowledge management*, páginas 1873–1876. ACM.

[Prasad2010] Prasad, Suhaas. 2010. Micro-blogging sentiment analysis using bayesian classification methods. En *Technical Report*. Stanford University.

[Reed y Rowe2004] Reed, Chris y Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979.

[Reimers et al.2019] Reimers, Nils, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, y Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

[Rousseeuw y Kaufman1990] Rousseeuw, Peter J y L Kaufman. 1990. Finding groups in data. *Hoboken: Wiley Online Library*, 1.

[Santus et al.2018] Santus, Enrico, Hongmin Wang, Emmanuele Chersoni, y Yue Zhang. 2018. A rank-based similarity metric for word embeddings. En *ACL*.

[Stab y Gurevych2014a] Stab, Christian y Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. En *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, páginas 1501–1510.

[Stab y Gurevych2014b] Stab, Christian y Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. En *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 46–56.

[Stab y Gurevych2017] Stab, Christian y Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

[Stab et al.2018] Stab, Christian, Tristan Miller, Benjamin Schiller, Pranav Rai, y Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 3664–3674.

[Tai, Socher, y Manning2015] Tai, Kai Sheng, Richard Socher, y Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

[Teufel y others1999] Teufel, Simone y others. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. tesis, Citeseer.

[Walton1996] Walton, Douglas N. 1996. *Argumentation schemes for presumptive reasoning*. Psychology Press.

[Wang y Cardie2016] Wang, Lu y Claire Cardie. 2016. A piece of my mind: A sentiment analysis approach for online dispute detection. *arXiv preprint arXiv:1606.05704*.

[Ward Jr1963] Ward Jr, Joe H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

[Zhao y Karypis2002] Zhao, Ying y George Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. En *Proceedings of the eleventh international conference on Information and knowledge management*, páginas 515–524.

Apéndice A

Features: Indicadores Léxicos

La tabla C1 muestra los indicadores léxicos utilizados presentados en (Stab y Gurevych, 2017). La lista incluye 22 indicadores de avance, 33 indicadores de retroceso, 48 indicadores de tesis y 10 de refutación.

Categoría	***Indicadores***
Avance (24)	"As a result", "As the consequence", "Because", "Clearly", "Consequently", "Considering this subject", "Furthermore", "Hence", "leading to the consequence", "so", "So", "taking account on this fact", "That is the reason why", "The reason is that", "Therefore", "therefore", "This means that", "This shows that", "This will result", "Thus", "thus", "Thus, it is clearly seen that", "Thus, it is seen", "Thus, the example shows"
Retroceso (33)	"Additionally", "As a matter of fact", "because", "Besides", "due to", "Finally", "First of all", "Firstly", "for example", "For example", "For instance", "for instance", "Furthermore", "has proved it", "In addition", "In addition to this", "In the first place", "is due to the fact that", "It should also be noted", "Moreover", "On one hand", "On the one hand", "On the other hand", "One of the main reasons", "Secondly", "Similarly", "since", "Since", "So", "The reason", "To begin with", "To offer an instance", "What is more"
Tesis (48)	"All in all", "All things considered", "As far as I am concerned", "Based on some reasons", "by analyzing both the views", "considering both the previous fact", "Finally", "For the reasons mentioned above", "From explanation above", "From this point of view", "I agree that", "I agree with", "I agree with the statement that", "I believe", "I believe that", "I do not agree with this statement", "I firmly believe that", "I highly advocate that", "I highly recommend", "I strongly believe that", "I think that", "I think the view is", "I totally agree", "I totally agree to this opinion", "I would have to argue that", "I would reaffirm my position that", "In conclusion", "in conclusion", "in my opinion", "In my opinion", "In my personal point of view", "in my point of view", "In my point of view", "In summary", "In the light of the facts outlined above", "it can be said that", "it is clear that", "it seems to me that", "my deep conviction", "My sentiments", "Overall", "Personally", "the above explanations and example shows that", "This, however", "To conclude", "To my way of thinking", "To sum up", "Ultimately"
Refutación (10)	"Admittedly", "although", "Although", "besides these advantages", "but", "But", "Even though", "even though", "However", "Otherwise"

Tabla C1: Lista de indicadores léxicos utilizada en *Argument Mining*, utilizada para extraer *features* para construir nuestro clasificador genérico en la sección 3.2.2.1

Todo list

Categoría	***Indicadores***
Evolutivo (48)	"feel", "change", "better", "worse", "improve", "worsen", 'side', "effect", "life", "deteriorate", "good", "bad", "results", "aggravate", "more", "less", "able", "worth", "keep", "decision", 'finally", "decision", "increase", "decrease", "begin", "start", "stop", 'switch", "help", "turn", "still", "save", "many", "safe", "rest", "helpful", "med", "normal", "excess", "other", "medicine", "dose", "improvement", "additive", "dosage", "allow", "something", "anything", "nothing", "easy", "hard", "all"
Temporal (33)	"day", "hour", "minute", "week", "month", "year", "hr", "min", "mth", "yr", "time", "once", "twice", "single", "before", "now", "after", "afterwards", "daily", "weekly", "monthly", "more", "less", "never", "ever", "again", "far", "soon", "since", "ago", "normally", "often", "regularly"
Personal (18)	"doctor", "dr", "physician", "psychiatrist", "psychologist", "therapist", "wife", "husband", "son", "daughter", "friend", "colleague", "boyfriend", "girlfriend", "couple", "partner", "family", "relationship"

Tabla C2: Lista de indicadores léxicos extraídos a partir del análisis del corpus *Drug Review Dataset*, como se describe en la sección 3.2.2.2