Combinación de LLMs basados en Transformers con información socio-demográfica para detectar contenido sexista en redes sociales

Trabajo Fin de Máster

Jacobo J. Pedrosa Marín

Trabajo de investigación para el

Máster en Lenguajes y Sistemas Informáticos

Universidad Nacional de Educación a Distancia

Dirigido por

Prof. Dr. D. Jorge Amando Carrillo de Albornoz Cuadrado y
Prof. Dra. Dña. Laura Plaza Morales

Febrero 2024

Agradecimientos

Quiero dedicar este trabajo ante todo a mi mujer, por toda la paciencia que ha tenido (y tiene) conmigo, sin su apoyo esto no habría sido posible. También a Diego, por hacerme sonreír.

A Jorge Carrillo y Laura Plaza por su ayuda, su disposición, constante orientación y confianza depositada en mí y en este trabajo, así como a todo el equipo docente de la Universidad y a mis compañeros en el departamento de lenguajes y sistemas.

A mi madre allá donde estés y a mi padre, sin ellos, no habría llegado hasta aquí.

Gracias a Sergio y a Silvia, por enseñarme que la familia está por encima de los lazos de sangre.

Finalmente, gracias a todos los amigos y familiares que me han apoyado durante estos años, gracias.

Resumen

Este trabajo se desarrolla en el marco de la edición de 2023 de EXIST, que consta de una serie de encuentros científicos y desafíos colaborativos destinados a la identificación del sexismo en plataformas de redes sociales. Su propósito abarca desde la detección de misoginia manifiesta hasta la identificación de comportamientos sexistas, sutiles y tácitos. La tercera entrega de este desafío conjunto se realizará como parte de un laboratorio en la conferencia CLEF 2023. En esta edición de EXIST, además del tema central de identificación de sexismo, las tareas se abordan desde la perspectiva de aprendizaje con desacuerdos *(learning with disagreements)*, donde cada instancia del conjunto de datos aportado, está asociada a seis etiquetas, las cuales se derivan de las anotaciones proporcionadas por anotadores pertenecientes a seis cohortes distintas (en función de género y edad).

A lo largo de este trabajo, se repasan tanto el estado del arte en torno a la detección toxicidad en internet de forma general, y más concretamente en torno a la identificación de sexismo, y se repasan las estrategias más comunes en cuanto al tratamiento de desacuerdo entre anotadores. Tras este análisis inicial, se plantean tres propuestas para la tarea 1 y otras 3 para la tarea 2, donde se alcanza la segunda posición en la métrica *soft-soft* en el contexto monolingüe español y la tercera posición en el contexto bilingüe. Además, la propuesta realizada, es la única en plantear un sistema basado en la información socio-demográfica de los anotadores, creando un modelo para cada cohorte para calcular la distribución final de probabilidades. Todo ello se recoge en un artículo científico que es enviado a la competición.

Finalmente, se extraen las conclusiones de los resultados obtenidos y se proponen cuáles podrían ser las siguientes líneas futuras de investigación tanto para la detección de sexismo como para la gestión de tareas con desacuerdo.

Abstract

This work is developed within the framework of the 2023 edition of EXIST, which consists of a series of scientific gatherings and collaborative challenges aimed at identifying sexism on social media platforms. Its purpose spans from detecting overt misogyny to identifying subtle and implicit sexist behaviors. The third installment of this joint challenge will be conducted as part of a laboratory at the CLEF 2023 conference.

In this edition of EXIST, in addition to the central theme of sexism identification, tasks are approached from the perspective of learning with disagreements, where each instance of the provided dataset is associated with six labels derived from annotations provided by annotators belonging to six different cohorts (based on gender and age).

Throughout this work, the state-of-the-art in detecting internet toxicity in general, and more specifically in identifying sexism, is reviewed, as well as the most common strategies for handling annotator disagreements. Following this initial analysis, three proposals are presented for task 1 and another 3 for task 2, where the second position is achieved in the soft-soft metric in the Spanish monolingual context and the third position in the bilingual context. Additionally, the proposal made is the only one to propose a system based on the socio-demographic information of the annotators, creating a model for each cohort to calculate the final probability distribution. All this is documented in a scientific paper that is submitted to the competition.

Finally, conclusions are drawn from the obtained results, and suggestions are made for potential future research directions for both sexism detection and managing tasks with disagreements.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

En este capítulo se exponen al lector la motivación propuesta y objetivos de este trabajo final de máster en Tecnologías del Lenguaje. Además, se introduce al lector en la problemática a tratar y se incluye la estructura de este trabajo con la intención de facilitar tanto la lectura como la comprensión del mismo.

1.1. Motivación

El aumento exponencial de la interacción en línea a través de redes sociales y plataformas digitales ha abierto nuevas fronteras en la comunicación global, proporcionando un espacio para el intercambio de ideas, la colaboración y la expresión personal. Esta proliferación de las redes sociales y plataformas digitales ha transformado la forma en que nos comunicamos y compartimos información a escala global. Por otro lado, este crecimiento exponencial también ha exacerbado la proliferación del contenido perjudicial, incluyendo el discurso tóxico en los espacios virtuales, planteando serias preocupaciones sobre la seguridad y el bienestar de los usuarios en estos entornos.

Este discurso perjudicial se puede presentar de diferentes formas, ya sea acoso, violencia, extremismo, etc. Y en su efecto en el resto de usuarios, se hace notar en diferentes formas, desde el abandono de la propia red social hasta trastornos o traumas psicológicos. Por otro lado, si bien este tipo de comportamiento ya existía anteriormente a las redes sociales, estas lo facilitan eliminando fronteras como el horario, la distancia, la memoria, etc.

El planteamiento de este trabajo se centra en el impacto negativo del

discurso sexista en las redes sociales, que no solo afecta a las personas directamente implicadas, sino que también perpetúa normas sociales dañinas y promueve la desigualdad de género en la sociedad en su conjunto. Ante esta problemática, surge la necesidad imperiosa de desarrollar soluciones innovadoras que puedan identificar y mitigar eficazmente el contenido sexista en línea, contribuyendo así a la creación de entornos digitales más seguros, inclusivos y respetuosos.

Además de esta motivación inherente de contribuir a una necesidad social, como es conseguir un entorno online seguro para todos, me planteo este proyecto como un desafío personal para trabajar con las tecnologías más recientes en el estado del arte del procesamiento del lenguaje natural. Entender y aplicar estas tecnologías no solo me permitirá abordar eficazmente los desafíos específicos del proyecto, sino también adquirir habilidades valiosas y estar a la vanguardia en un campo en constante evolución. La motivación para utilizar transformers en una tarea de clasificación y clasificación multi-clase radica en su capacidad demostrada para abordar eficazmente una amplia gama de problemas de procesamiento del lenguaje natural (NLP). El enfoque de transformers es altamente adaptable y escalable, lo que lo hace adecuado para diversos conjuntos de datos y contextos de aplicación. La posibilidad de explorar y aplicar esta tecnología en tareas de clasificación y clasificación multi-clase representa una oportunidad emocionante para ampliar mi experiencia y conocimiento en NLP, así como para obtener resultados más precisos y sólidos en estas áreas específicas.

Por otro lado, desde el punto de vista de anotación de un conjunto de datos, es importante tener en cuenta que no tiene por qué haber consenso entre los anotadores respecto a un tema concreto como es la identificación de sexismo. El planteamiento de tecnología con información socio-demográfica para entender el punto de vista de diferentes personas respecto a un mismo tema, puede suponer un avance no solo a nivel científico, sino a nivel social, permitiendo identificar si un contenido puede ser ofensivo para una persona en concreto, puede contribuir en gran medida a reducir este tipo de comportamientos que en muchos casos.

1.2. Definición del problema

En los últimos años, las redes sociales han supuesto un gran impacto tanto en el ámbito de la comunicación como directamente en la sociedad. Los servicios de microblogging y las redes sociales como Twitter y Facebook han demostrado su potencial para transmitir ideas, pensamientos y conocimiento entre sus usuarios. Estas plataformas ofrecen unos canales de comunicación sin límites, y es precisamente esta ausencia de límites junto con la anonimidad de los usuarios que permite, aparte de los mismos, utilizar estos canales para atacar a individuos concretos (ciberbullying) o a grupos en general (lenguaje de odio, racismo, misoginia...).

Tal es la necesidad de tomar medidas en este campo que en febrero de 2015 el CEO de Twitter de aquel momento, declaró que este tipo de lenguaje genera abandono de sus usuarios [1] y en 2016, las grandes plataformas de Internet como Facebook, Google, Twitter y Microsoft firmaron un acuerdo con la Unión Europea con el fin de eliminar todo el contenido de odio de sus plataformas en 24 horas [2]. De esta manera, se puede observar que en la política de uso de estas plataformas especifican que este tipo de lenguaje no es admitido por las mismas:

> "Conducta que incita al odio: no puedes atacar directamente a otras personas ni atacarlas o amenazarlas directamente por motivo de su raza, origen étnico, origen nacional, pertenencia a una casta, orientación sexual, género, identidad de género, afiliación religiosa, edad, discapacidad o enfermedad grave" [3].

En el estudio (Duggan, 2017) realizado en el *Pew Research Center*[4] se ponen cifras a este problema: 4 de cada 10 adultos en USA experimentan algún tipo de acoso online y, en la mayoría de los casos, este acoso se centra en características personales, el aspecto físico o el género de la víctima, tal como se puede observar en la imagen 1.1.

[1] https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the

[2] https://techcrunch.com/2016/05/31/facebook-twitter-youtube-and-microsoft-agree-to-remove-hate-speech-across-the-eu/?guccounter=1

[3] https://help.twitter.com/es/rules-and-policies/twitter-rules

[4] El Pew Research Center es un centro de datos independiente que realiza estudios públicos sobre los problemas, actitudes y tendencias en USA y el Mundo.

[5] https://www.pewresearch.org/internet/2017/07/11/experiencing-online-harassment/

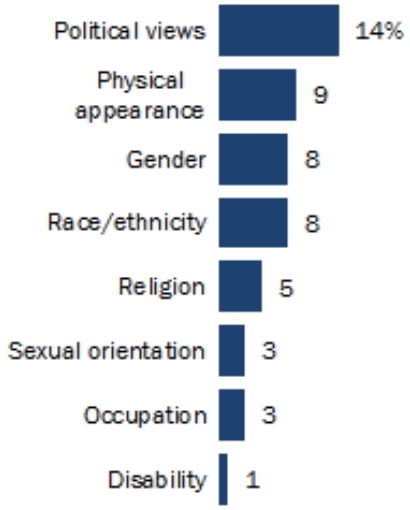

Figura 1.1: Tipos de acoso online: En esta imagen se listan diferentes comportamientos en línea de los que han sido víctimas los encuestados para el informe Pew Reseach Center[5].

En este estudio se valora el efecto que tiene este acoso en la propia víctima a partir de la percepción que se tiene sobre esta situación en función de diferentes características y se observa que, en general, las mujeres consideran que es un problema serio, que no se le da suficiente importancia y se debería hacer.

En el estudio (Vogels, 2021) posterior realizado en el mismo centro, pese a que se determina que el acoso online ha frenado su crecimiento respecto a 2014, se sigue observando una gran diferencia con respecto a otros tipos de acoso que puede recibir un ciudadano adulto en USA. En este segundo estudio, se profundiza en los perfiles de las víctimas de estos tipos de acoso, donde se puede observar que destacan, por un lado, las personas entre 18 y 29 años y sobre todo aquellas personas con una inclinación sexual no heterosexual. Finalmente y, sobre todo, de las conclusiones de este estudio, cabe destacar que pese a que los hombres tienen más tendencia a recibir algún tipo de acoso online, el 70 % de las mujeres opinan que es un problema importante.

Para las plataformas sociales en internet, la detección de contenido tóxico, no es un tema trivial, dada la cantidad de usuarios y el contenido que se

[6]https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

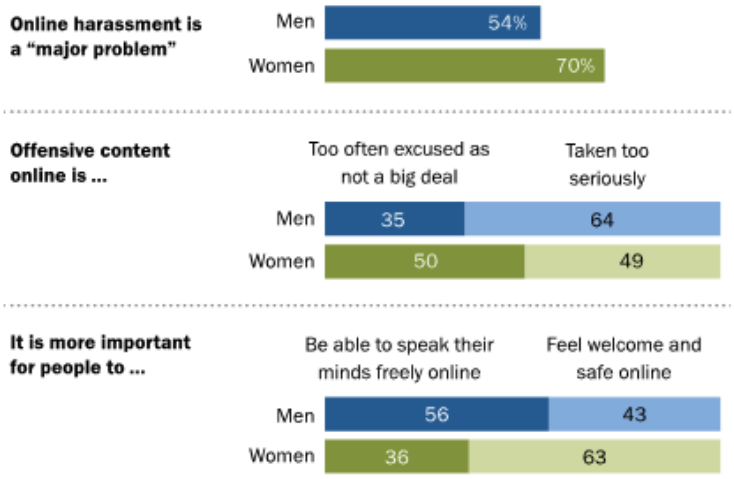

Figura 1.2: Percepción del acoso en Internet según el género: en estas gráficas comparativas se muestra que si bien tanto hombres como mujeres perciben el acoso en línea como un problema grave, la percepción respecto al uso y contenido online varía entre géneros[6].

crea por segundo en estas plataformas. El control y la moderación de estos contenidos son dos de los problemas principales a los que se tienen que hacer frente. Es por ello que año a año se desarrollan, por parte tanto de estas empresas, entidades independientes e incluso entidades públicas, sistemas que permitan analizar automáticamente el contenido publicado y detectar aquellas publicaciones que, entre otras, tengan base en algún tipo de discurso de odio, ya sea a una persona o grupo. Estas herramientas van desde tecnologías capaces de gestionar conjuntos de datos cada vez más grandes a otras que minimicen el error en la detección de este tipo de contenidos.

1.3. Propuesta y objetivos

El principal objetivo tras este trabajo es, por tanto, abordar esta problemática del sexismo en redes sociales utilizando técnicas avanzadas de procesamiento del lenguaje basadas en *Large Language Models* LLMs y transformers, con la intención de aprovechar su capacidad para comprender y analizar el lenguaje humano en diferentes contextos.

Como segundo objetivo, se pretende entender la dinámica y estrategias en general que se utilizan para tratar el problema de toxicidad, y más concre-

tamente sexismo en redes sociales. Para ello se pretende abordar y repasar el estado del arte relacionado con estas temáticas.

Por otro lado, tanto en el campo del sexismo, en el de la toxicidad en general, como en un sentido aún más amplio en cualquier campo donde pueda existir diversidad de opinión, el desacuerdo es un aspecto intrínseco del mismo. Es por ello que en este trabajo se pretende abordar también el paradigma del desacuerdo entre anotadores de una manera transversal, repasando las diferentes aproximaciones y planteamientos existentes en el estado del arte del momento.

Finalmente, para el desarrollo de este trabajo, se propone trabajar en el marco de la competición EXIST 2023 en el foro CLEF del mismo año. En esta competición se proponen 3 tareas basadas en la detección y clasificación de sexismo en tuits tanto en español como en inglés:

- **Tarea 1 - Detección de sexismo:** la primera tarea se trata de una clasificación binaria donde el sistema propuesto debe determinar si un tuit es o no sexista.
- **Tarea 2 - Clasificación de la intención de la fuente:** en esta tarea se pretende categorizar el contenido del tuit en función de la intención del autor permitiendo determinar el papel que juega este dentro de la red social en la emisión y difusión de mensajes sexistas. Las posibles categorías en esta tarea son:
  - i) Mensaje directamente sexista (direct sexist message)
  - ii) Reporte de mensaje sexista (reported sexist message)
  - iii) Mensaje de juicio de valor (judgemental message)
- **Tarea 3 - Categorización de sexismo:** en esta tercera tarea, cada tuit se categoriza en una o varias de las siguientes categorías que determinan el foco del mensaje sexista:
  - (i) Ideología y desigualdad (ideological and inequality)
  - (ii) Estereotipado y dominación (stereotyping and dominance)
  - (iii) Objetificación (objectification)
  - (iv) Violencia Sexual (sexual violence)
  - (v) Misoginia y violencia no-sexual (misogyny and non-sexual violence)

Para cada una de estas tareas se debe generar una respuesta en formato de etiqueta única *hard* y una segunda respuesta con la probabilidad de cada etiqueta posible *soft*.

En el marco de esta competición, se proporciona un conjunto de datos de entrenamiento de 6.400 tuits en Español e Inglés (3.200 por idioma), uno de desarrollo que contiene 500 tuits por idioma y finalmente un conjunto de test de 1.000 tuits por idioma. Con el fin de obtener diferentes perspectivas y reducir en la medida de lo posible el sesgo en fase de anotación, cada tuit es anotado por seis anotadores de seis cohortes diferentes basadas en género (únicamente se considera hombre/mujer debido a razones de disponibilidad) y edad (18-22 años, 23-45 años y +46 años). En la publicación (Plaza et al., 2023) se amplía la información sobre el conjunto de datos.

En desarrollo de este trabajo, y para enfrentar las 2 primeras tareas expuestas en la competición, se implementan diferentes modelos basados en aumento de datos y la tecnología Transformer con los que se realizan 3 propuestas para cada tarea y de entre todas las métricas de evaluación, se llega a obtener en una de ellas la segunda posición.

1.4. Estructura del documento

Esta sección presenta la estructura definida en este trabajo.

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. Caso de Estudio propuesto: EXIST 2023. En este capítulo se describe la edición de EXIST 2023 en la que se ha participado en sus dos primeras tareas como parte de este trabajo final de máster

Capítulo 4. Combinación de LLMs basados en Transformers con información socio-demográfica para detectar contenido sexis-

ta en redes sociales. En este capítulo se describe en profundidad el sistema/método o caso de estudio propuesto.

Capítulo 5. 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 mostrar los resultados de los sistemas presentados a la campaña de evaluación.

Capítulo 6. 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 se presenta una descripción general del estado del arte en torno a la detección de diferentes tipos de contenido tóxico en redes sociales, centrándose a continuación en el sexismo. Posteriormente, se repasa la trayectoria de la competición EXIST repasando sus ediciones anteriores e introduciendo la edición EXIST 2023 que embarca este trabajo. Finalmente, se muestran las técnicas y estrategias más utilizadas para resolver problemas de*Learning With Disagreement*.

2.1. Toxicidad en Redes Sociales

Una de las características principales de la llamada web 2.0, es la direccionalidad de la comunicación, teniendo los usuarios una participación activa en la creación del contenido en la misma. Uno de los ejemplos más representativos de esta tendencia sería Wikipedia[1] donde, a partir del formato "Wiki", el contenido es creado y gestionado por la propia comunidad. Otros ejemplos en los que se muestra la importancia de esta participación de los usuarios serían el auge de los comentarios en publicaciones en blogs o webs de noticias, valoraciones de un producto en e-commerce o, de forma más general, las redes sociales, donde cada usuario además puede unirse o crear una comunidad con personas con los mismos intereses, todo ello de manera global y en entornos donde se puede interactuar con millones de usuarios de todo el planeta.

Esta facilidad de comunicación, y la libertad de publicar de forma anóni-

[1] https://wikipedia.com

ma, también ha generado un aumento del alcance y la magnitud del contenido perjudicial o tóxico disponible en la web, que incluye desinformación, teorías conspirativas, extremismo, acoso, violencia y otras formas de material socialmente tóxico (Sheth, Shalin, y Kursuncu, 2022). Esto es especialmente cierto en las redes sociales, pero también en los otros entornos online, como los anteriormente mencionados.

Si bien la mayoría de estos incidentes potencialmente perjudiciales, como el acoso escolar o los discursos de odio, existían antes de la llegada de Internet, el alcance y la extensión de la red les da un poder y una influencia sin precedentes para afectar las vidas de miles de millones de personas(Kumar et al., 2018a) y (Duggan, 2017). Estas situaciones no solo generan aflicción mental y psicológica entre los usuarios de la web, sino que además han forzado a individuos a desactivar sus cuentas y, en casos extremos, incluso han desencadenado suicidios. Por consiguiente, los actos de agresión y el comportamiento verbal inapropiado no se han mantenido como simples molestias menores, sino que afecta a un gran número de personas. Es por esto que resulta crucial implementar medidas preventivas para lidiar con la conducta abusiva y agresiva en línea (Kumar et al., 2018a).

En respuesta a esta problemática, la detección y clasificación de lenguaje tóxico en Internet es un creciente campo de interés tanto para empresas, como usuarios e investigadores, con el fin de poder moderar este tipo de conductas en plataformas sociales. Ejemplos de estas medidas a destacar serían en el ámbito académico, el auge de tanto de workshops, tareas y competiciones (Waseem et al., 2017a)(Fišer et al., 2018)(Roberts et al., 2019)(Akiwowo et al., 2020)(Kumar et al., 2018b)(Kumar et al., 2020)(Kumar et al., 2022) como de publicaciones en conferencias de prestigio (**?**) y en el ámbito empresarial, proyectos como el clasificador de contenido abusivo desarrollado por el equipo de Yahoo (Nobata et al., 2016), el proyecto **Perspective**, desarrollado por Google y Jigsaw (Hosseini et al., 2017) y en el que se plantea una herramienta para detección de contenido tóxico que pueda ser utilizada por cualquier plataforma a través de una API, y por su parte Twitter ha lanzado en los últimos años diferentes herramientas como “Unmention” que permite a las personas autoexcluirse de una conversación en la que han sido mencionados [2] o “Twitter Circle” que es un sistema de reportes renovado

[2]https://blog.twitter.com/common-thread/en/topics/stories/2022/dont-at-me-protecting-your-peace-with-unmention

con controles de respuestas más avanzados con el fin de promover un entorno más seguro y respetuoso.

2.2. Detección y clasificación de toxicidad

Los investigadores aún no han alcanzado un consenso sobre lo que constituyen términos como toxicidad o acoso en línea (Risch y Krestel, 2020), pero en general se describe como el tipo de comunicación que es perjudicial, hiriente o de naturaleza negativa que puedan llegar a causar daño emocional o psicológico a quienes lo reciben. Además, la dificultad de analizar el lenguaje natural hace que la detección y clasificación de contenido tóxico en cualquiera de sus formas esté lejos de ser una tarea trivial que pueda realizarse mediante una simple detección de patrones o palabras clave. Tanto si se trata de detección automática como de un trabajo realizado por humanos, esta tarea requiere de un conocimiento general del mundo y del ámbito del mensaje. Sin embargo, algunos de los aspectos que definen contenidos de este tipo pueden ser **insultos**, **discriminación**, **difamación** o **alegaciones** y/o **definiciones no verificables** (Risch y Krestel, 2020).

Dado el contexto en el que se encuentran estos contenidos, algunos de los aspectos que pueden afectar la anotación de un mensaje tóxico son, la **ofuscación de palabras clave** *"ni99er"*, **aparición constante de nuevos insultos** que no se hayan incluido previamente en una lista negra, **el contexto** en el que se pronuncia una frase, ya sea geográfico, político o social, y finalmente, la intencionalidad del emisor, donde textos claramente tóxicos, pueden haber sido enunciados en tono de ironía o sarcasmo lo que afecta a los resultados obtenidos (Nobata et al., 2016).

Dada esta problemática, se pueden encontrar en la bibliografía diferentes iniciativas orientadas a la detección de contenido tóxico. En este sentido, destacan los talleres de **TRAC** (Trolling, Aggression and Cyberbulling) (Kumar et al., 2018b)(Kumar et al., 2020) y (Kumar et al., 2022), que se centran en detectar los diferentes formatos de agresión en línea entre los que destacan el "trolling" y el "ciberacoso", siendo esta su tarea transversal en sus diferentes ediciones. A partir de la segunda edición, se añade además la detección de misoginia, a partir de la identificación del género a quien va dirigida la publicación (Kumar et al., 2020).

Algunas de las tareas más incipientes en este sentido son: Hateval (Basile

et al., 2019a) y Offenseval (Zampieri et al., 2019) y (Zampieri et al., 2020) en la edición de Semeval de 2019. La primera, centrada en la detección de lenguaje de odio orientado a dos grupos concretos, los inmigrantes y las mujeres. Para ello se propone una primera tarea de clasificación binaria en la que, a partir de un conjunto de datos extraído de Twitter, los sistemas deben predecir si un tuit muestra o no odio hacia el grupo de los inmigrantes o de las mujeres. Offenseval, por otro lado, se centra en la detección y clasificación de contenidos ofensivos. A raíz de las dos ediciones de 2019 y 2020, se crean dos conjuntos de datos. El primero, **OLID** (Offensive Language Identification Dataset), contiene 14.200 tuits anotados en inglés utilizando un modelo de 3 niveles: (A) si el contenido es ofensivo, (B) categorización del tipo de lenguaje ofensivo y (C) objetivo a quien va dirigida la ofensa. El segundo, **SOLID** (Semi-Supervised Offensive Language Identification Dataset), está compuesto por cerca de nueve millones de tuits etiquetados de forma semisupervisada. La tabla 2.1 muestra un resumen estadístico para los dos conjuntos de datos para cada nivel de taxonomía.

Level	Label	OLID		SOLID	
		Train	Test	Train	Test
A	OFF	4,640	240	1,448,861	3,002
	NOT	9,460	620	7,640	2,991
B	TIN	4,089	213	149,550	1,546
	UNT	551	27	39,424	1,451
C	IND	2,507	100	120,330	1,055
	GRP	1,152	78	22,176	349
	OTH	430	35	7,043	140

Tabla 2.1: Resumen de los conjuntos de datos creados en el proyecto OffensEval en sus dos ediciones. OLID en 2019 y SOLID en 2020.

Cabe destacar la tarea *"Profiling Hate Speech Spreaders on Twitter"* (Rangel et al., 2021) en la edición del CLEF de 2021 en la que los sistemas deben determinar dado un hilo de Twitter si su autor difunde odio. Para ello, se crea un conjunto de datos bilingüe, Español-Inglés, en el que cada tuit está anotado según 3 clasificaciones binarias: si el autor difunde odio o no, en caso afirmativo, si el odio va dirigido a un grupo o una persona en concreto y finalmente si el autor muestra o no un comportamiento agresivo además de odio.

Otra tarea que cabe destacar en esta línea, es **Detoxis** *"DEtection of TOXicity in comments In Spanish"* (Taulé Delor et al., 2021), cuya intención es la detección de toxicidad en comentarios publicados en español en respues-

ta a diferentes artículos de noticias en línea relacionados con la inmigración. Para ello se plantean dos subtareas: la primera, de detección de toxicidad, es una tarea de clasificación binaria que consiste en clasificar el contenido de un comentario como tóxico (tóxico=sí) o no tóxico (tóxico=no). La segunda, de detección del nivel de toxicidad, es una tarea de clasificación, en la que el objetivo es identificar el nivel de toxicidad de un comentario (0= no tóxico; 1= ligeramente tóxico; 2= tóxico y 3: muy tóxico).

Por otro lado, se pueden encontrar diferentes publicaciones en las que se proponen técnicas de clasificación diferentes. Por ejemplo, (Clarke y Grieve, 2017), utilizan el análisis multidimensional (MDA) (Biber, 1988)(Biber, 1989) para extraer las dimensiones más relevantes de la variación lingüística mediante un análisis factorial, que son interpretadas en función de las características lingüísticas y los textos individuales que están más fuertemente asociados a cada dimensión.

En (Wiegand, Siegel, y Ruppenhofer, 2018) dentro del Workshop GermEval de la edición de 2018, se definen dos subtareas: la clasificación binaria de sexismo en tuits y la clasificación más detallada sobre el tipo de ofensa que se encuentra en cada contenido: (i) OTHER, cuando el tuit no es ofensivo; (ii) PROFANITY, para tuits con palabras profanas que no pretenden ofender a nadie; (iii) INSULT, cuando el tuit claramente quiere ofender a alguien; y finalmente, (iv) ABUSE, cuando el tuit no pretende insultar a alguien en concreto, pero tiene una clara intención de degradar a alguien o a un colectivo.

En (Waseem et al., 2017b) se propone un esquema bidimensional de lenguaje abusivo con dos dimensiones: "generalizado/dirigido" y "explícito/implícito". En (Risch y Krestel, 2020) se propone una categorización para comentarios tóxicos inspirada en anotaciones realizadas para diferentes conjuntos de datos:

- **Lenguaje obsceno/blasfemias**, que consisten en palabras malsonantes y posibles variaciones de las mismas y suelen detectarse utilizando listas negras.
- **Insultos**, declaraciones groseras u ofensivas que afectan a un individuo o a un grupo.
- **Amenazas**, comentarios que anuncian o abogan por infligir castigo,

dolor, lesiones o daño a uno mismo o a otros.

- **Discurso de odio/odio a la identidad**, dirigido exclusivamente a grupos definidos por religión, orientación sexual, etnia, género u otros identificadores sociales. Se atribuyen atributos negativos al grupo como si estos fueran universalmente válidos.

- **Otros comentarios** que no se ajustan a ninguna de las cuatro clases anteriores, pero que probablemente harían que otros usuarios abandonen una discusión, se consideran "tóxicos" sin más especificaciones, en esta categoría se pueden contemplar el trolling, spam, etc...

En los comentarios clasificados como de **discurso de odio** (o *Hate Speech*) destacan sus formas de **racismo** y **sexismo**, siendo esta última la categoría en la que se centra este trabajo.

2.3. Sexismo en Redes Sociales

Dentro del contenido tóxico en general, destaca el sexismo como forma de odio a un grupo en concreto, tal es el alcance de este problema que Amnistía Internacional publicó en 2020 un informe [3] en el que se describe la plataforma Twitter como un "lugar tóxico" para las mujeres(Rodríguez-Sánchez, Carrillo-de Albornoz, y Plaza, 2020). En la misma línea, el congreso de los EEUU, ha reclamado a Facebook más acciones de protección a las mujeres frente a ataques de este tipo[4]. Existen a su vez numerosos estudios como (Duggan, 2017) o (Vogels, 2022) en los que se demuestra que las mujeres, sobre todo en edad escolar, es el público más frecuente para cualquier tipo de ataque en Internet. En la imagen 2.1 se observa cómo, de todos los públicos evaluados, son las mujeres de entre 15 y 17 años las que sufren más ciberbullying.

Como se afirma en (Rodríguez-Sánchez, Carrillo-de Albornoz, y Plaza, 2020), el sexismo se encuentra con frecuencia en diversas formas en las redes sociales e incluye una amplia gama de comportamientos (como estereotipos, problemas ideológicos, violencia sexual, etc.), y es por ello que la detección y clasificación de contenido sexista es una tarea cada vez más frecuente en

[3] https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1-1/

[4] https://www.reuters.com/article/us-facebook-women-politics-idUSKCN2522KK/

[5] https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/

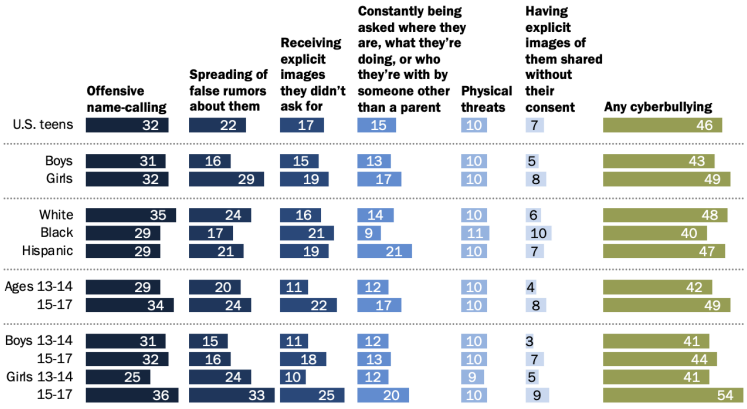

Figura 2.1: Análisis de ciberbullying: en esta imagen se muestran diferentes tipos de ataque online que reciben los jovenes en USA separando rasgos sociales como género, edad y raza. Se resalta que las mujeres de entre 15 y 17 años son las que reciben más ataques en todas las categorías respecto a los otros estratos definidos[5].

congresos y workshops dentro del ámbito de Procesamiento del Lenguaje Natural.

Una de las tareas más incipientes en este sentido es la segunda tarea de la edición de 2020 de TRAC (Kumar et al., 2020), en la que tras haber realizado una identificación de agresión en publicaciones de Twitter, el participante puede tratar de identificar el género de a quién va dirigida dicha agresión. De este modo, si la víctima es una mujer se determina dicho contenido como misógino.

Merecen especial atención las tareas de AMI *"Automatic Misogyny Identification"* (Anzovino, Fersini, y Rosso, 2018). Constan de dos subtareas: identificación de misoginia en tuits y reconocimiento del objetivo: (i)**activo**, cuando se refiere a una persona individual; (ii) **pasivo**, cuando el objetivo es genérico o un grupo, y el tipo de los mensajes misóginos; (i) **descrédito**, cuando el objetivo es únicamente insultar; (ii) **estereotipo y objetivación**, cuando se pretende definir a la mujer por ideas básicas y preconcebidas como si fuera un objeto; (iii) **dominación**, cuando se pretende establecer la superioridad del hombre frente a la mujer; (iv) **descarrilamiento**, cuando se pretende justificar el abuso sobre la mujer; (v) **acoso sexual y violencia**, donde el mensaje es de índole directamente sexual, insinuaciones o violencia.

En la segunda edición de AMI(Fersini, Nozza, y Rosso, 2020), donde las dos subtareas se centran primero en la detección de agresividad y misoginia

en genera, clasificando cada contenido en (i) **misógino**, si el texto expresa odio hacia las mujeres en particular; (ii) **no misógino**, si el texto no expresa ninguna forma de odio hacia las mujeres; (iii) **agresivo**, en caso de que el contenido del texto tenga algún tipo de agresividad ya sea implícita o explicita; y (iv) **no agresivo**, si no se da ninguna de las anteriores. En la segunda subtarea de esta edición, se plantea trabajar en el problema de la introducción de sesgo a partir de *términos de identidad* que están fuertemente relacionados con la misoginia, sin ser exclusivos de este tipo de mensajes. Por ejemplo, la palabra "mujer" suele estar incluida en un texto misógino, pero no todas las publicaciones con esta palabra se pueden clasificar como tal. Para ello, se implementa un conjunto de datos sintético para test basándose en la publicación de (Basile et al., 2019b), en la que se recogen todos aquellos términos relacionados con identidad y son substituidos por $< identity\_term >$ de manera que se elimina la identidad del mensaje pero no su objetivo.

Posteriormente, la tercera edición de AMI, denominada MAMI *"Multimedia Automatic Misogyny Identification"* al añadir elementos multimedia como memes en el conjunto de datos. En esta, las tareas se centran en la detección de memes misóginos en la web y en la clasificación del tipo de misoginia en (Fersini et al., 2022).

2.4. EXIST (sEXism Identification in Social neTworks)

La detección de toxicidad en redes sociales es, por tanto, un foco de atención en la comunidad académica y en la empresarial tanto por sus perjuicios sociales como económicos. También se ha visto que el contenido tóxico abarca muchas formas, entre las que destaca el que centra su atención en las mujeres.

A diferencia de otros formatos de toxicidad en redes sociales, el sexismo abarca muchas formas, desde las más sutiles al odio más acérrimo (Rodríguez-Sánchez, Carrillo-de Albornoz, y Plaza, 2020). Una de las principales dificultades en su detección es que la propia naturaleza de este tipo de comentarios puede ser difícil de discernir, sobre todo según el perfil de la persona. Frases como *"hay que amar y respetar a las mujeres y siempre tratarlas como si fueran de cristal"*, *"No trates de entender a las mujeres, es*

imposible", o *"haz feliz a tu mujer regalándole una lavadora"* son mensajes que de por sí no representan aparentemente ningún tipo de lenguaje de odio, pero en su significado se está menospreciando a la mujer definiéndola como un género más débil que el masculino. Tal como se muestra en el estudio (Swim et al., 2001), este tipo de mensajes pueden afectar incluso más que aquellos más cercanos al odio dado que no solo muestran una expresión respecto a la mujer, sino que están difundiendo una manera de pensar y una cultura sobre la misma que afecta a como se ve a la mujer y al comportamiento que "debería" tener tanto en la vida personal como profesional.

Esta sección se centra en la competición EXIST[6] y en las tres ediciones de la misma: las dos precedentes (Rodríguez-Sánchez et al., 2021) y (Rodríguez-Sánchez et al., 2022) que se realizan dentro del el marco de IberLeft *"Iberian Languages Evaluation Forum"* dentro del SEPLN *"Sociedad Española para el Procesamiento del Lenguaje Natural"* y la que abarca este trabajo (Plaza et al., 2023) que se desarrolla dentro del marco de CLEF 2023 *"Conference and Labs of the Evaluation Forum"*.

EXIST como competición se basa en que la detección del sexismo en general. En sus dos primeras ediciones, se proponen dos subtareas: identificación de sexismo, donde los sistemas deben predecir si un tuit es o no sexista ("SEXIST"/"NOT SEXIST") y categorización del tipo de sexismo en una de las siguientes cinco categorías:

- **Ideological and inequality (Ideología y desigualdad):** textos en los que se rechaza la existencia de desigualdad de género, el movimiento feminista y/o presenta a los hombres como víctimas de la opresión de género.

  > *"Perdonar lo no sólo es posible, es absolutamente necesario. Estoy harta de las boludas que acusan de todos los males de la humanidad a los hombres. Muerte al feminismo"*

- **Stereotyping and dominance (Estereotipado y dominación):** donde la publicación expresa ideas falsas sobre la mujer en las que se sugiere que es más adecuada para cumplir determinados roles (ama de casa, esposa, cuidadora, amorosa, sumisa, etc...) o a la inversa, no adecuada para determinadas tareas (conducir, trabajo duro, etc...) e

[6] http://nlp.uned.es/exist2023/

incluso donde se afirma que el hombre es superior a la mujer en algún sentido.

> *"Desde que nacemos nos enseñan a diferenciar a la perfección de lo que está bien o lo que está mal.¿Sacar buenas notas? Perfecto.¿Que se comporta como una señorita? Bien.¿Que quiere llevar una falda corta?muy puta.¿Delgada? Éxito ¿Gorda?Perdedora¿Que llora?Infantil..."*

- **Objectification (objetivación):** textos en los que se presenta a la mujer como un objeto, independientemente de su dignidad y/o individualidad. Generalmente, se asumen o describen cualidades físicas que las mujeres deben tener para cumplir con roles tradicionales preestablecidos, como cánones de belleza, hipersexualización, etc...

  > *"Busco mujer que viva en Torrevieja o cerca de Alicante para follar"*

- **Sexual violence (Violencia sexual):** contenidos en los que se hacen referencias explícitamente sexuales, solicitud de favores de este tipo, acoso e incluso mención a violación o agresión sexual.

  > *"Te quiero follar hasta romperte las piernas. Eres una niña mala, una niña mala, ¿Te gusta eso? Sí, te encanta. Maldita puta ofrecida. Vamos, gímeme, sigue gimiendo. No te detengas, pequeña zorra ofrecida. ¿Te gusta esto? Sí, te fascina, te excita, mh. /rugirle."*

- **Misogyny and non-sexual violence (Misoginia y violencia no-sexual)**: textos en los que se expresa directa o indirectamente odio hacia las mujeres.

  > *"Lo que es indignante es que una frígida como tú se lo lleve doblado y los autónomos nos morimos de hambre."*

Pese a que el planteamiento inicial de las dos primeras ediciones de EXIST es muy similar, se plantea una diferencia principal en la anotación del conjunto de datos de test que se utiliza en la segunda respecto a la primera. En los siguientes sub-apartados, se explica como se definen los conjuntos de datos utilizados y una breve aproximación a las diferentes propuestas utilizadas en ambas ediciones.

2.4.1. EXIST 2021

Esta primera edición de EXIST en 2021 se define como la primera tarea compartida de detección de sexismo en redes sociales cuyo objetivo principal es la identificación y clasificación de sexismo en el sentido amplio, desde el mensaje sexista más explicito a aquellas expresiones más sutiles como las mencionadas anteriormente. Para ello, se crea un dataset a partir de publicaciones en español e inglés de Twitter y de la red social sin censura Gab[7] en el que se incluyen todas estas formas de mensajes sexistas.

Para la creación de este conjunto de datos, primero se hace una selección de expresiones y palabras clave, tanto en español como en inglés, utilizados en comentarios que afectan de algún modo a la mujer. Esta selección es revisada y filtrada por Trinidad Donoso and Miriam Comet de tal manera que se obtiene un conjunto resultante de 116 términos de búsqueda en español y 109 en inglés. A partir de estos términos de búsqueda, mediante el uso de la API de Twitter, se descargan 545.717 tuits en español y 662.895 en inglés de fechas comprendidas entre el 1 de diciembre de 2020 y 28 de febrero de 2021. Para asegurar un correcto balanceo entre los contenidos obtenidos para cada palabra clave, se descartan aquellos términos con menos de 60 tuits, quedando finalmente un conjunto de 91 términos semilla para español y 93 para inglés.

Para las publicaciones de Gab.com, se descargan el volcado más reciente de esta plataforma desde pushshift[8] y buscando aquellas publicaciones que contuvieran los términos de búsqueda seleccionados de tal manera que se obtienen 1.583 gabs en español y 1.356.266 gabs en inglés.

A partir de estos contenidos, se define un set de entrenamiento de 4500 tuits para cada idioma y otro de prueba de 2000 tuits también por idioma. Finalmente, en el conjunto de prueba se añaden 500 gabs por cada idioma, de tal manera que se pueda analizar las diferencias entre redes sociales con y sin control.

Para el etiquetado de estos conjuntos de datos, inicialmente, se desarrolló una guía de anotación en inglés y español en la que se proporcionaba una explicación clara de cada etiqueta junto con varios ejemplos. Para evaluar la calidad de dicha guía de anotación, se etiquetaron 50 tuits en español, obteniendo un índice kappa de 0.58 para la tarea 1 y 0.45 para la tarea 2.

[7] https://gab.com
[8] https://pushshift.io/

Estos resultados indicaron un acuerdo moderado que coincidió con el hecho de que la detección de sexismo desde una perspectiva amplia no es simple, dando a entender que el sexismo es aún más subjetivo que la misoginia o que el discurso de odio hacia las mujeres, por lo que el proceso de etiquetado tiende a ser más difícil.

Los resultados de este experimento se utilizaron para modificar la guía de anotación, a partir de de la cual, se realizó un experimento de anotación utilizando Amazon Mechanical Turk [9] (MTurk). Para ello, se definió un gold standard de un subconjunto de datos de 100 tuis en español y 100 en ingles y fue etiquetado por dos expertos (un hombre y una mujer con 2 años de experiencia en clasificación de sexismo), cuyos casos de desacuerdo fueron resueltos por un tercer contribuyente experimentado. Por otro lado, cada tuit de este subconjunto fue anotado por 5 anotadores de crowdsourcing, siguiendo la guía definida, y el resultado final se definió por el voto mayoritario de los anotadores de crowdsourcing, el cual fue comparado con la etiqueta seleccionada por los expertos. Como se puede observar en la tabla 2.2 de resultados de este experimento,estos indican un acuerdo sustancial por lo que los anotadores realizaron correctamente la tarea, por lo que se mantuvo este sistema para anotar el conjunto de datos total revisando manualmente aquellas instancias con 3 votos en una clase para la tarea 1 y con desacuerdo en la tarea 2.

	Tarea 1		Tarea 2	
	Kappa	% Acuerdo	Kappa	% acuerdo
Español	0.74	0.87	0.57	0.71
Inglés	0.62	0.83	0.49	0.72

Tabla 2.2: Análisis del acuerdo EXIST 2021

En esta edición de EXIST, para la primera subtarea de las propuestas, se presentaron 31 equipos, 70 propuestas (runs) de los cuales finalmente 26 obtienen un "Accuracy" superior a la línea base (baseline) definida inicialmente y solo cinco equipos quedan por debajo de esta media. A continuación se detallan las cinco propuestas con mejores resultados de ambas tareas:

- La propuesta de (de Paula, da Silva, y Schlicht, 2021) del grupo **AI-UPV** presenta un sistema basado en el aumento de datos mediante

[9] https://www.mturk.com

traducción de los tuits en español a inglés y viceversa. A continuación se crean diferentes modelos basados en BERT para los tuits en inglés, BETO para español y mBERT como modelo multiidioma. A partir de estos modelos, se crean sistemas que combinan los resultados obtenidos por cada modelo para obtener una respuesta basada en diferentes estrategias como voto mayoritario o valor estándar. Esta propuesta es la que obtuvo mejores resultados en las dos tareas con un *Accuracy de 0.7804* y *F1 de 0.7802* en la primera tarea y un *Accuracy de 0.6577* y un *F1 de 0.5787* en la segunda.

- El sistema desarrollado por el equipo **SINAI TL**(Plaza-del Arco et al., 2021) se basa en el uso de otros conjuntos de datos de tareas relacionadas con la identificación de sexismo en Twitter (InterTASS 2019, EmoEvent, HatEval y MEX-A3T), a partir de los cuales se crea un sistema multitarea donde el objetivo es aprender diferentes tareas a la vez para mejorar el rendimiento de cada tarea; en este caso el sistema se basa en BETO para los contenidos en español y BERT para inglés. Los resultados de este equipo para las dos tareas fueron un *Accuracy de 0.7800* y *F1 de 0.7797* como su mejor resultado en la primera tarea y un *Accuracy de 0.6527* y un *F1 de 0.5667* en la segunda.

- En el caso del sistema propuesto por el equipo **AIT FHSTP 2**(Schütz et al.,), se desarrolla mediante el uso de dos modelos multiidioma, el primero en multilingual BERT y el segundo en XLM-R. Para las dos tareas, los mejores resultados se obtienen con los modelos basados en XLM-R. Los resultados de este equipo para las dos tareas fueron un *Accuracy de 0.7754* y *F1 de 0.7752* como su mejor resultado en la primera tarea y un *Accuracy de 0.6445* y un *F1 de 0.5589* en la segunda.

- El equipo **Multiaztertest**, únicamente presenta propuestas para la primera tarea, donde el mejor resultado se basa en el uso de BERT y BETO para creación de modelos, obteniendo un *Accuracy de 0.774* y *F1 de 0.7731*(Rodríguez-Sánchez et al., 2021).

- El equipo **NLP UNED**, plantea la comparación de diferentes tipos de sistema (mono-tarea y multitarea) basados en Transformers. En la

primera tarea obtienen un mejor resultado con un sistema mono-tarea basado en XLM-T con un *Accuracy de 0.772* y *F1 de 0.7702* y en la segunda con un sistema multitarea basado en XLM-R que obtiene *Accuracy de 0.06232* y *F1 de 0.5509*(Rodríguez-Sánchez et al., 2021).

Aparte de estos resultados destacados, en la edición de 2021 de EXIST, la mayoría de las contribuciones para ambas tareas se basaron en modelos basados en transformadores para la clasificación. De los 23 equipos participantes, 14 equipos utilizaron BERT, un modelo transformador ampliamente utilizado, como base para sus soluciones, 10 equipos emplearon BETO, una versión de BERT que está entrenada en texto en español y finalmente, 5 equipos aprovecharon XLM-R, una variante multilingüe de RoBERTa que admite varios idiomas, incluido el español (Rodríguez-Sánchez et al., 2021).

Es interesante destacar diferencia entre los mejores y los peores sistemas en esta edición de EXIST 2021, donde en la tarea 1 los mejores 10 sistemas se basaron en tecnología transformers. En la tarea 2 los resultados varían entre 0.5787 y 0.1069 en la métrica F1 y los resultados más bajos los obtuvieron los equipos que utilizaron técnicas tradicionales de aprendizaje automático, como SVM (Support Vector Machine) y RF (Random Forest).

2.4.2. EXIST 2022

La tarea compartida, propuesta en EXIST 2022, se basa en las mismas subtareas que en la edición de 2021, identificación y categorización de sexismo en tuits, manteniendo el formato del conjunto de datos y las mismas categorías para la segunda subtarea utiliza además, el mismo conjunto de datos de EXIST 2021 (entrenamiento + test) como conjunto de entrenamiento en el que se incluyen tanto los contenidos de Twitter como los de Gab que eran exclusivos del conjunto de test en la edición anterior. Sin embargo, en esta edición se crea un nuevo conjunto de test con contenido exclusivo de Twitter, y con un cambio significativo: la anotación se realiza por voto mayoritario a partir de las anotaciones realizadas por seis expertos (tres hombres y 3 mujeres para prevenir el sesgo por género) entrenados específicamente para esta tarea (Rodríguez-Sánchez et al., 2022).

En esta edición de EXIST, se presentaron 60 grupos de los que 19 envían propuestas para la primera tarea y 15 para la segunda. Cabe destacar que todos los equipos participantes, excepto uno, utilizaron soluciones basadas en transformers. Entre estas soluciones, ocho equipos utilizaron BERT, cinco

equipos utilizaron BETO y cuatro equipos utilizaron RoBERTa. La adopción generalizada de modelos basados en transformers en ambas ediciones de EXIST resalta su eficacia para abordar los desafíos asociados con la detección de sexismo en las redes sociales (Rodríguez-Sánchez et al., 2022). A continuación se detallan algunas de las propuestas que mejores resultados han obtenido en una o ambas tareas.

- **Avacaondata 1 - task 1&2**: la propuesta realizada por el equipo Avacondata(Tamayo y Bueno,) destaca por obtener la primera posición en ambas tareas. El sistema se centra en una combinación de diferentes modelos basados en transformers, por un lado BERTweet-large (Nguyen, Vu, y Nguyen, 2020), RoBERTa (Liu et al., 2019a) y DeBERTa v3 (He, Gao, y Chen, 2021) para ingés y por otro, BETO (Cañete et al., 2020a), BERTIN (Javier et al., 2022), MarIA-base (Gutiérrez-Fandiño et al.,) y RoBERTuito (Pérez et al., 2022) para español. Todos estos modelos se entrenan en dos fases, en la primera se utiliza un set de validación para optimización de hyperparámetros y a continuación se entrenan los modelos con el conjunto de datos entero. Los resultados obtenidos por este equipo en las dos tareas son: un *Accuracy de 0.7996* y *F1 de 0.7978* en la primera tarea y un *Accuracy de 0.7013* y un *F1 de 0.5106* en la segunda.

- **CIMATCOLMEX - task 1**(Vaca-Serrano, 2022): pese a que solo presentó propuesta para la primera tarea, obtuvo un resultado muy cercano al anterior equipo mediante una combinación de diez modelos basados en RoBERTuito (Pérez et al., 2022) y diez basados en BERT (Devlin et al., 2018a) entrenados cada uno de ellos de manera independiente utilizando diferentes semillas. El resultado que obtuvo este equipo para la primera tarea es un *Accuracy de 0.7949* y un *F1 de 0.7940*.

- **I2C 1 - task 1**: La propuesta (Álvarez et al., 2022) se basa principalmente en aumentar el dataset inicial traduciendo los contenidos de español a inglés y viceversa. A continuación obtiene los resultados a partir de una combinación de tres modelos entrenados con este dataset RoBERTa (Liu et al., 2019a), BETO (Cañete et al., 2020a) y SiEBERT (Hartmann et al., 2023). Los resultados obtenidos por este equipo en las dos tareas son: un *Accuracy de 0.07883* y *F1 de 0.7880*

en la primera tarea y un *Accuracy de 0.6465* y un *F1 de 0.4700* en la segunda.

- **Multiaztertest - tasks 1&2**: La propuesta (Bengoetxea, 2022) se desarrolla en base a RoBERTa (Liu et al., 2019a) para los tuits en inglés y BETO (Cañete et al., 2020a) para los tuits en español. Los resultados de este equipo para las dos tareas fueron un *Accuracy de 0.7836* y *F1 de 0.7830* como su mejor resultado en la primera tarea y un *Accuracy de 0.6786* y un *F1 de 0.4706* en la segunda.

- **ELiRF-VRAIN 3 - task2**: La propuesta de (Ahuir, 2022) realiza un aumento de datos mediante dos estrategias. La primera y más básica consiste en traducción del inglés al español y viceversa. La segunda consiste en enmascarar determinados tokens aleatoriamente. Este conjunto de datos aumentado se utiliza para entrenar, por un lado, cinco modelos diferentes para español basados en XLM-R (Conneau et al., 2019), RoBERTa (Liu et al., 2019a) y otras tres variantes de BERT(Devlin et al., 2018a) y, por otro lado, otros cinco modelos para los contenidos en inglés basados en XLM-R(Conneau et al., 2019), RoBERTa(Liu et al., 2019a), BERT(Devlin et al., 2018a), hateBERT(Caselli et al., 2020) y ALBERT(Lan et al., 2019). Los resultados de este equipo para las dos tareas fueron un *Accuracy de 0.7694* y *F1 de 0.7686* como su mejor resultado en la primera tarea y un *Accuracy de 0.7042* y un *F1 de 0.4991* en la segunda, siendo este además el tercer mejor resultado tras los dos del equipo **Avacondata**.

A partir de los resultados de las dos ediciones anteriores de EXIST se observa que es destacable el uso que se hace de modelos basados en tecnología Transformer como BERT y modelos derivados con los que se obtienen los mejores resultados en la mayoría de los casos. Así mismo, destacan estrategias de combinación de diferentes modelos y el aumento de datos del conjunto de entrenamiento a partir de la traducción bidireccional entre español e inglés de los diferentes tuits de los que está conformado el mismo.

2.4.3. EXIST 2023

Los cambios más significativos en esta edición de EXIST 2023 son la introducción de una nueva tarea y la creación de un nuevo conjunto de

datos, elaborado siguiendo la metodología de *"Learning With Disagreement (LeWiDi)*. A las dos tareas de las ediciones anteriores (identificación de sexismo y categorización de sexismo) se añade una nueva tarea: identificación de la intención del autor. En concreto consiste en clasificar cada tuit en una de las siguientes categorías:

- **Directo *(DIRECT)***: donde la intención del autor del mensaje es compartir o difundir un mensaje sexista por sí mismo o con intención de inducir al sexismo.

  > *"Bueno por lo que vi, a las mujeres hay que presumirlas y darles atención perfecto anotado. Lo peor de todo es que nos gorrean igual"*

- **Reportado *(REPORTED)***: la intención es reportar y/o compartir una situación de sexismo sufrida por una o varias mujeres/es ya sea en primera o tercera persona.

  > *"#PorSiNoLoViste #Espectaculos #Acoso #MeToo La #top-model británica Emily Ratajkowski denunció que fue acosada sexualmente por el cantante Robin Thicke, durante el rodaje del videoclip de BlurredLines DiariodeMexico"*

- **Juicio de valor *(JUDGEMENTAL)***: la intención es juzgar o criticar comportamientos o situaciones sexistas con la intención de condenarlos.

  > *"espero que lo haya querido decir es que gracias a la pornografia y la cultura de la violacion, para muchas las relaciones sexuales pueden llegar a ser violentas por la idea de dominación/sumisión que esta tan erotizado que se use fuerza física contra nosotras https://t.co/6TXZpQiVxy"*

Esta tarea pasa a ser la tarea 2 de EXIST 2023, mientras que la tarea 3 cincide con la tarea 2 de las dos ediciones anteriores EXIST 2021 y EXIST 2022 con la única diferencia de que en EXIST 2023 pasa a ser una tarea de clasificación multi etiqueta; es decir cada tuit puede asociarse a más de una categoría de sexismo.

Pero este no es el único cambio de esta edición de EXIST 2023, para esta edición se crea un nuevo conjunto de datos en el que, además de trabajar con

nuevos tuits, se introduce el concepto del *desacuerdo entre anotadores*. Bajo el paradigma de *"Learning with disagreements"*; el conjunto no proporciona un *gold standard* con el que el participante tiene que entrenar sus modelos, sino que se ofrece la perspectiva de los diferentes anotadores (6 en este caso). Para ello, han participado en el proceso de anotación 1,065 anotadores a través de la plataforma **Prolific**[10] que se han dividido en 6 estratos en función de su edad (18-22/23-45/+46) y género (masculino/femenino).


```
{
  "100030": {
    "id_EXIST": "100030",
    "lang": "es",
    "tweet": "@Quora Is a great example of this, they don't care if I have being
      supplanted, harassed or else, even with evidence the problem was me. Por
      supuesto en Quora español. Hay 1 tipo con cientos de cuentas ofensivas y
      sigue ahí. Y el clonador de cuentas... https://t.co/cRNZ0VNnUn https://t.co
      /oQnHzyOrsN",
    "number_annotators": 6,
    "annotators": [ "Annotator_109","Annotator_110","Annotator_111"
      ,"Annotator_112","Annotator_113","Annotator_114"],
    "gender_annotators": ["F","F","F","M","M","M"],
    "age_annotators": ["18-22","23-45","46+","46+","23-45","18-22"],
    "labels_task1": ["YES","YES","NO","YES","NO","NO"],
    "labels_task2": ["REPORTED","REPORTED","-","REPORTED","-","-"],
    "labels_task3": [
      ["STEREOTYPING-DOMINANCE","OBJECTIFICATION","SEXUAL-VIOLENCE","MISOGYNY-NON
        -SEXUAL-VIOLENCE"],
      ["SEXUAL-VIOLENCE"],
      ["-"],
      ["OBJECTIFICATION"],
      ["-"],
      ["-"]
    ],
    "split": "TRAIN_ES"
  }
}
```


Figura 2.2: Ejemplo de anotación de un tuit en el que se puede observar que para cada tarea hay diferentes anotaciones, una por cada cohorte.

Cohorte	Genero	Edad
1	Hombre	18-22
2	Mujer	18-22
3	Hombre	23-45
4	Mujer	23-45
5	Hombre	+46
6	Mujer	+46

Tabla 2.3: Diferentes cohortes o grupos de anotadores en EXIST 2023

El conjunto de datos se crea a partir de 8.000.000 de tuits en Inglés y Español descargados entre el 1 y el 30 de setiembre de 2022 utilizando las

[10]https://prolific.co

mismas palabras clave que en las otras ediciones y para mitigar el sesgo que pueda haber se han eliminado aquellos tuits relacionados con una palabra clave que a su vez tenga menos de 60 tuits. Finalmente, se trabaja con conjuntos de palabras clave o palabras semilla de 183 para español y 163 para inglés(Plaza et al., 2023).

Por último, el tercer cambio significativo de esta nueva edición de EXIST 2023, se centra en la evaluación de los sistemas participantes. Es necesario tener en cuenta que al incluir desacuerdo en cada una de las instancias, la evaluación de los sistemas debe también tener en cuenta el desacuerdo. Es por ello que en esta edición la evaluación no se puede hacer con métricas como *Accuracy* o *FMeasure* como en las ediciones anteriores, sino que se debe acudir a un planteamiento diferente. Para ello se deben considerar los dos extremos de la evaluación, **groud tryth** o verdad fundamental que es el valor final que se extrae del desacuerdo. Esta puede plantearse de un modo *hard* (estricto) definiendo una única categoría para cada instancia a partir de las anotaciones, por ejemplo, mediante voto mayoritario, o en un escenario *soft* no estricto, en el que se conserve además de la categoría el desacuerdo en formato de distribución de probabilidad. En este escenario cabe tener en cuenta que las dos primeras tareas al tratarse de problemas mono-etiqueta, la suma de las probabilidades ha de ser 1, en cambio, en la tercera tarea, al ser un problema multi etiqueta, la suma de las probabilidades puede ser mayor que 1. En el otro extremo de la evaluación se encuentra la salida de los sistemas, que también pueden dar una respuesta en formato de una única etiqueta en una configuración **hard** o predecir la probabilidad de cada etiqueta en una configuración **soft**.

Es por ello que, para las diferentes tareas de esta edición de EXIST 2023, se consideran tres tipos de evaluación:

- **Hard-hard (hard system output & hard ground truth):** en este tipo de evaluación se intenta comparar una única etiqueta propuesta por cada sistema una única etiqueta en el *gold standard* y dado que en el conjunto original. Para ello, y dado que en el conjunto original cada tuit està asociado a seis etiquetas diferentes, en la primera tarea se considera la clase que haya sido anotada por más de tres anotadores; en la segunda por más de dos y finalmente en la tercera, que es de naturaleza multi etiqueta, aquella clase con más de una anotación entre los diferentes anotadores. En aquellos casos en los que no se encuentre

una clase mayoritaria, la instancia directamente no se considerará para la evaluación.

- **Hard-soft (hard system output & soft ground truth):** en este esquema se pretende comparar la salida *hard* de los sistemas frente a las probabilidades de cada categoría obtenida de la distribución de clases resultante de la anotación. Para ello, y dado que no existe ninguna métrica actual que se ajuste a problemas de clasificación jerárquica de múltiples etiquetas en un escenario de múltiples anotaciones, se utiliza una extensión de ICM (ICM-soft) creada específicamente para esta edición de EXIST 2023 y que se explica en la sección 5.1.1
- **Soft-soft (soft system output & soft ground truth):** esta evaluación permite comparar la respuesta de los sistemas dada como una distribución de probabilidad frente a las probabilidades obtenidas para cada clase de la anotación humana.

2.5. Learning With Disagreement

Tal como se ha comentado en la introducción de este trabajo, uno de los aspectos más relevantes de la detección de contenido tóxico en redes sociales es el perfil sociocultural de tanto el emisor, el receptor y el público general que puede leer esta publicación, ya que el mismo mensaje se puede interpretar de diferentes maneras según su contexto.

Además, en las publicaciones de las ediciones de 2021 y 2022 de EXIST, uno de los puntos destacables en la creación del conjunto de datos en general de la primera competición y del conjunto de test en la segunda, es el desacuerdo existente entre los anotadores, incluso cuando se trata de expertos debidamente entrenados. Tal es la importancia de este punto, que en la segunda edición de EXIST, tras la creación de una guía en inglés y español donde se provee de una clara explicación de cada etiqueta junto con diversos ejemplos, se realiza un experimento para anotar el 20 % del conjunto de datos de pruebas en el que se obtiene un resultado de 0.387 kappa para la primera tarea y 0.336 en la segunda, lo cual indica un pobre acuerdo y obliga a revisar la guía de anotación creada (Rodríguez-Sánchez et al., 2022).

Este problema es común en procesos de anotación donde trabaje más de un anotador (ya sean desarrolladores, investigadores o expertos independientes), y viene dado por el propio sesgo individual que toda persona tiene

imbuido, ya sea por sus características socio demográficas, profesionales, culturales, etc.

En la publicación de (Poesio y Artstein, 2005), se identifican hasta tres limitaciones principales en el estado del arte en cuanto a anotación de relaciones anafóricas: (i) falta de estudio sistemático del acuerdo en la anotación del problema en cuestión (anáforas en la literatura), (ii) comprensión limitada del grado de acuerdo sobre las referencias a entidades abstractas como en la deixis discursiva y (iii) que una expresión pueda tener más de una interpretación en el contexto de su ocurrencia. Estas 3 limitaciones se pueden extrapolar a otros ámbitos como al que pertenece este trabajo, (i) la clasificación de un contenido tóxico puede en gran medida determinarse con las características socioculturales del que la lee, es importante, por tanto, tener en cuenta el emisor, receptor y contexto de un mensaje para poder determinar correctamente la tipología a la que pertenece dicho contenido, (ii) es imprescindible, por tanto, entender el contexto en el que se da un mensaje, y finalmente (iii) tener en cuenta que un mensaje puede tener más de una interpretación, por lo que es importante trabajar con conjuntos de datos que puedan ofrecer este desacuerdo.

Existe, además, diferente literatura respecto al papel del desacuerdo en el análisis y detección del discurso de odio, en (Aroyo et al., 2019) se aborda el problema de la toxicidad online teniendo en cuenta el sesgo subjetivo. determinando que la manera que tenemos de percibir algo como tóxico se ve influido por el contexto previo o actual, tal como cultura, antecedentes, experiencias, educación, etc. En esta misma publicación, se clasifican dos razones para el sesgo: i) la subjetividad inherente al tema, que se da en aquellos temas en los que la valoración depende directamente de la experiencia personal de cada uno y donde directamente dos individuos pueden simplemente opinar diferente; ii) la subjetividad condicionada por la ambigüedad de la tarea ya sea porque exista ambigüedad en la explicación o que las entidades sean ambiguas por sí mismas.

En (Davani et al., 2023) se evalúa el papel de los estereotipos sociales en la detección automática del discurso del odio en inglés examinando el impacto de los estereotipos sociales en los resultados de la anotación, los conjuntos de datos anotados y los clasificadores del discurso de odio. Para ello se investiga el impacto de los estereotipos de un conjunto de anotadores novatos en su anotación de un conjunto de datos basado en *hate speech* donde

se detecta que cuanto más alto es el nivel de competencia de quien recibe un mensaje, se asignan más etiquetas de discurso de odio, se muestra una mayor tendencia a identificar el discurso de odio y se reduce el desacuerdo entre anotadores.

En la publicación (Basile et al., 2021), se defiende que es erróneo plantear tareas basándose en una verdad fundamental en el campo del Procesamiento del Lenguaje Natural, dado que en muchos casos existe más de una respuesta correcta e incluso en los casos en los que se dé esta posibilidad, el desacuerdo entre anotadores afecta en el momento de determinar el *gold standard*. Por tanto, es necesario primero capturar el desacuerdo entre anotadores y definir nuevos sistemas de evaluación que lo permitan.

Por tanto, el desacuerdo es un valor intrínseco a cualquier conjunto de datos en los que intercede el valor humano en el momento de clasificar los contenidos, sobre todo, en un mundo global donde el contexto y la opinión de un individuo puede variar en función de su situación geográfica, edad, género, educación, etc. Es por tanto, imprescindible encontrar vías, permitan estudiar y/o el desacuerdo. Para ello, en este apartado se van a repasar tanto la tarea basada en *"Learning With Disagreements"* de la edición de SemEval de 2021 (Uma et al., 2021a) como las principales estrategias en el estado del arte en cuanto a tratamiento del desacuerdo entre anotadores.

2.5.1. SemEval 2021 - Learning with Disagreements

En la edición de SemEval de 2021[11], se celebra la tarea 12 de *Learning with Disagreements*, en la que se propone un entorno de pruebas unificado para estudiar el desacuerdo en las áreas de *Natural Language Processing* y *Computer Vission*. Para ello se basan en conjuntos de datos publicados en ambas áreas, todos ellos planteados en la perspectiva de proveer múltiples etiquetas para cada instancia que evidencien diferentes características entre los anotadores y los procedimientos de obtención de los datos, y por otro lado, un tamaño suficiente para entrenar modelos en el estado del arte. Para esta tarea, se proponen cinco conjuntos de datos que incluyen el desacuerdo original de los anotadores: i) (Gimpel et al., 2011), en el que cada token de este conjunto de datos es etiquetado por entre 5 y 177 anotadores con una media de 16,38 anotadores por token. Para la tarea propuesta se seleccionan 8.3K, 3K y 3.1K tokens para los subconjuntos de entrenamiento, desarrollo

[11] https://semeval.github.io/SemEval2021/

y prueba respectivamente; ii)(Poesio et al., 2019) es un corpus creado en el marco del proyecto DALÍ [12] que lleva desde 2008 haciendo investigaciones en torno al desacuerdo dentro de NLP a gran escala. Para ello se basan en el desarrollo de plataformas gamificadas y/o juegos online mediante los que recopilar grandes cantidades de datos [13] relacionados con la detección de anáforas en textos. Para esta tarea en concreto, se ha utilizado una versión simplificada del corpus de PDIS con instancias etiquetadas de forma binaria **Discurse New** o **Discurse Old** dependiendo de si la entidad referenciada ha sido o mencionada anteriormente; iii)**The humour dataset**, esta subtarea se basa en el corpus creado por (Simpson et al., 2019) que está compuesto por textos cortos humorísticos como chistes, bromas o juegos de palabras y también textos no humorísticos como proverbios o aforismos. todos ellos emparejados y etiquetados por 5 anotadores indicando cuál de los dos textos en cada par es más humorístico; iv)**The LabelMe corpus**, esta tarea esta basada en uno de los conjuntos de datos más populares en Computer Vision, el LabelMe Dataset(Russell et al., 2007), que esta conformado por imágenes de exterior distribuidas en 8 categorías (autopista, interior de ciudad, edificios altos, calle, bosque, costa, montaña o espacios abiertos. Posteriormente Rodrigues y Pereira en 2018 (Rodrigues y Pereira, 2018) utilizan Mechanical Turk para obtener una media de 2.5 anotaciones por imagen de 59 anotadores para 10K imágenes en este conjunto; v)**CIFA-10h** (Krizhevsky, Hinton, y others, 2009) es un conjunto formado por 60K imágenes distribuidas entre 10 categorías avión, coche, pájaro, gato, perro, ciervo, rana, caballo barco o camión, a partir del cual (Peterson et al., 2019a), anotaron las imágenes utilizando Amazon Mechanical Turk con una media de 51.10 anotaciones por instancia.

2.5.2. Trabajar en conjuntos de datos sin acuerdo

En el campo de Inteligencia Artificial se suele trabajar con conjuntos de datos con instancias que tienen una única clase o interpretación conocida como **Gold Standard** o "estándar de oro". Sin embargo, este enfoque no logra captar las sutilezas del comportamiento humano en el que a menudo se encuentran desacuerdos y/o diferentes perspectivas. En las tareas en las que la subjetividad o ambigüedad forman parte inherente de la misma,

[12] http://dali.eecs.qmul.ac.uk/project
[13] https://anawiki.essex.ac.uk/phrasedetectives/

es imprescindible trabajar con conjuntos de datos que conserven este desacuerdo por parte de los anotadores. En este apartado se pretende repasar diferentes enfoques o estrategias existentes en el estado del arte para abordar esta problemática. Para ello se va a utilizar de base la publicación de (Uma et al., 2021b) que hace un muy buen primer análisis y catalogación de estos métodos. Según esta publicación, el estado del arte actual permite categorizar estas soluciones en cuatro categorías principales:

Agregación de juicios

Las estrategias basadas en agregación de juicios vienen encabezadas por la forma más popular de eliminar el desacuerdo entre anotadores, que es el voto mayoritario, donde para cada instancia se escoge la etiqueta con mayor número de votos entre los anotadores. Si bien es una estrategia simple y fácil de entender, no solo elimina el desacuerdo entre anotadores, sino que no tiene en cuenta la certeza de cada uno por separado, dando a todos los votos el mismo peso. Es por ello que dentro de esta categoría se pueden encontrar otras aproximaciones que pretenden gestionar esta limitación:

- **Métodos de agregación probabilística**: la agregación probabilística es uno de los métodos más comunes por su sencillez de implementación. Se incluye el llamado **voto mayoritario**, estrategia en la que el valor final para cada instancia es el que recibe más votos de entre los anotadores. Pero de esta forma se pierde el valor que puede aportar el desacuerdo en sí. Es por ello que en la publicación (Dawid y Skene, 1979) se plantea una de las primeras propuestas para abordar los problemas del voto mayoritario. Su enfoque calcula la probabilidad de una etiqueta para un elemento en función de la etiqueta observada, la etiqueta real, la prevalencia de las etiquetas y la probabilidad de que un anotador asigne una etiqueta en particular dada su etiqueta real, que se calcula a partir de las anotaciones de cada codificador. A partir de este planteamiento, se han desarrollado diferentes propuestas. En el trabajo de (Paun et al., 2018) se puede encontrar una visión general y una comparación de algunos de estos modelos.

- **El enfoque "CrowdTruth" para la agregación**: se basa en el trabajo de (Aroyo y Welty, 2015) en el que se desarrolla la idea de que el desacuerdo es una señal y no ruido, por lo que es preciso analizar-

lo y tenerlo en cuenta. Para ello, se planteó el desarrollo de nuevas métricas para evaluar la calidad de los anotadores, el acuerdo en las instancias (una medida de la dificultad del ítem) y el acuerdo en las etiquetas (Inel et al., 2014; Dumitrache et al., 2018; Dumitrache, Aroyo, y Welty, 2019), así como versiones revisadas de las métricas estándar de *precisión/recall/F-measure*.

- **Métodos de agregación basados en heurísticas y métricas**: los métodos de agregación basados en heurísticas son enfoques que utilizan reglas o técnicas basadas en la experiencia y el conocimiento experto para combinar o fusionar las respuestas o anotaciones de múltiples fuentes o anotadores en un solo resultado. Estos métodos no se basan en cálculos probabilísticos o estadísticos, sino que dependen de reglas predefinidas que buscan tomar decisiones informadas sobre cómo combinar las anotaciones. Pese a que se han propuesto diversos métodos de agregación basados en heurísticas (Hung et al., 2013; Sheshadri y Lease, 2013; Daniel et al., 2018), ninguno de estos ha demostrado superar al método de agregación probabilística propuesto por Dawid y Skene en cuanto a la obtención de "gold standard".

Filtrado de entidades complejas

En esta categoría se recogen diferentes publicaciones cuyo planteamiento principal es utilizar la información obtenida del propio desacuerdo para filtrar aquellas instancias que se consideren complejas, como por ejemplo aquellos los casos de gran desacuerdo entre los anotadores (Reidsma y op den Akker, 2008). En esta publicación se plantean 2 opciones para tratar el desacuerdo, la primera utilizando las instancias con gran acuerdo para entrenar el modelo y tratar las complejas como ruido, la segunda opción se centra en entrenar diferentes modelos uno por anotador para crear un sistema de voto que realice una predicción cuando los modelos converjan en una clase.

En las publicaciones (Beigman Klebanov, Beigman, y Diermeier, 2008), (Klebanov y Beigman, 2009), (Beigman y Beigman Klebanov, 2009), Beigman-Klebanov defiende que estas instancias con mayor desacuerdo deberían ser eliminadas, entendiendo que este tipo de instancias no se pueden considerar como buenos ejemplos del fenómeno que se está intentando enseñar al modelo. En (Beigman y Beigman Klebanov, 2009), se propone desarrollar modelos que permitan determinar la propia dificultad de una instancia en

función del desacuerdo que pueda gestionar el propio filtrado. Con lo que a partir de este planteamiento, se pueden encontrar diferentes aproximaciones que se centran ya no en el propio desacuerdo y/o el filtrado de instancias más problemáticas, sino en la dificultad de estas instancias a partir del desacuerdo. Algunos de estos acercamientos se centran en la creación de modelos de predicción de esta dificultad (Beigman y Beigman Klebanov, 2009), o incluso el análisis y detección de patrones en el propio desacuerdo (Reidsma y Carletta, 2008).

Entrenar un clasificador a partir del desacuerdo

Las publicaciones recogidas en esta categoría se basan en la asunción inicial de que no tiene por qué existir una única etiqueta en todos los casos y, por tanto, no se pueden basar en esta. El planteamiento, por tanto, se centra en intentar capturar el conocimiento global de los anotadores a partir de la distribución de etiquetas seleccionadas por los anotadores. Para ello se pueden sub clasificar las estrategias de esta categoría, a su vez, en tres opciones principales:

- **Repeated labeling** (Sheng, Provost, y Ipeirotis, 2008; Jamison y Gurevych, 2015): se incluyen los métodos que multiplican cada instancia por cada una de las anotaciones recibidas, pero de forma ponderada, asignando a cada etiqueta el peso de 1/x donde x es el número de anotadores, persiguiendo de este modo entrenar al modelo con el conocimiento colectivo de los anotadores y con las diferentes perspectivas de cada uno.

- **Soft loss functions**: estos métodos se centran en entrenar el modelo directamente de la distribución de probabilidad de las etiquetas a partir de las anotaciones como *"soft targets"* en la función de pérdida del modelo, ya sea mediante entropía cruzada o error cuadrático medio. En la publicación de (Peterson et al., 2019b), se comenta que el uso de entropía cruzada y *"soft labels"* y la distribución de probabilidad predicha por el modelo obtiene buenos resultados cuando se trata de generalizar sobre datos nuevos.

- **Learning from crowds**: en estos métodos se plantea desarrollar modelos a partir directamente del desacuerdo entre los anotadores. En la publicación de (Raykar et al., 2010), se hace uso de un algoritmo

de Esperanza-Maximización (EM) para entrenar un modelo de regresión/clasificación tanto con la precisión de cada anotador como con la etiqueta asignada a cada instancia. De esta manera, el algoritmo detecta directamente los mejores expertos y asigna mayor peso a sus predicciones. Finalmente, los resultados son ajustados por el algoritmo de EM que iterativamente modifica el propio *gold standard* con el fin de obtener mejor rendimiento. En (Albarqouni et al., 2016) se presenta una variación de esta propuesta basada en la agregación de datos directamente como parte del proceso de aprendizaje de una red convolucional (CNN) a través de una capa adicional llamada *AggNet*[14]. El sistema se basa principalmente en entrenar una serie de modelos basados en la misma arquitectura CNN con diferentes muestras y utilizar estos modelos para hacer unas predicciones previas. En este punto, los resultados obtenidos son enviados a diferentes anotadores cuyos resultados se envían nuevamente a la CNN existente. Este enfoque de múltiples capas se utiliza para garantizar anotaciones redundantes sobre las mismas instancias y así aumentar la solidez tanto de la agregación como de la clasificación. Finalmente, en esta línea de investigación cabe destacar la publicación de (Rodrigues y Pereira, 2018) ya comentada anteriormente en relación con el conjunto de datos creado y utilizado en la tarea propuesta de SemEval 2021. En su publicación, aparte de la propuesta de conjunto de datos, se plantea un sistema llamado **Deep learnign from crowds (DLC)** basado en la publicación de (Guan et al., 2017), donde se crea una capa de salida que da un resultado para cada anotador con la probabilidad de cada etiqueta. Este sistema muestra muy buenos resultados tanto en el campo de CV como en el de NLP es por ello que destaca en el ámbito general en el que se pueden encontrar diferentes anotadores y/o desacuerdo.

Uso de "hard labels" e información del desacuerdo

En esta categoría se recogen un conjunto de propuestas y publicaciones que se basan en la determinación de que si bien existe una etiqueta única o *gold standard* para cada instancia, también se reconoce la existencia de incertidumbre y estos modelos se pueden beneficiar del aprovechamiento de la información que puede aportar. En este sentido, la publicación de

[14]https://albarqouni.github.io/publication/albarqouni-2016-aggnet/

(Uma et al., 2021b) subdivide esta categoría a su vez en dos planteamientos principales:

- **Métodos que evalúan la incertidumbre entre anotadores** y la aprovechan para medir la pérdida asociada a cada instancia. Estos métodos vienen encabezados por la propuesta de (Plank, Hovy, y Søgaard, 2014) que, a partir de las anotaciones de dos anotadores sobre un conjunto de datos, mide la incertidumbre entre ambos anotadores, primero utilizando la métrica F1 y, posteriormente, calculando la probabilidad de confusión entre etiquetas. Otro método similar viene en la publicación de (Sharmanska et al., 2016), la que, al igual que algunas de las aproximaciones descritas en la categoría de filtrado de etiquetas complejas, se centra en medir la dificultad de determinar una etiqueta a partir de la confusión entre anotadores, pero en este caso en lugar de utilizar dicha información para filtrarlas, se añade al clasificador como medida de confianza, informando al modelo de cuánta importancia debe dar a cada instancia en el entrenamiento.

- **Métodos que entrenan modelos a partir de etiquetas y desacuerdo** como, por ejemplo, el presentado en (Lalor, Wu, y Yu, 2019) que se propone el entrenamiento de modelos utilizando etiquetas ("gold/hard labels") en una epoca, y la distribución de probabilidad ("soft labels") en la siguiente, o utilizar las primeras para entrenamiento y las segundas para "fine-tuning".

Capítulo 3

Caso de Estudio propuesto: EXIST 2023

En el capítulo 2 de este trabajo se introduce la competición EXIST a partir de las dos ediciones anteriores en 2021 y 2022. En ellas se plantean dos tareas de clasificación, la primera de clasificación binaria, donde se tiene que determinar si el contenido de un tuit es sexista o no y la segunda de clasificación multi-clase, donde el objetivo es determinar el tipo de sexismo. En el presente capítulo se describe la edición de EXIST 2023 en la que se ha participado en sus dos primeras tareas como parte de este trabajo final de máster.

3.1. EXIST 2023

Si bien las dos ediciones anteriores de EXIST se dan en el marco de IberLeft (Rodríguez-Sánchez et al., 2021; Rodríguez-Sánchez et al., 2022), el desarrollo de la tercera edición de esta competición se realiza en el marco del foro CLEF[1]. CLEF 2023 es la 14ª edición de la conferencia CLEF que se celebra desde el año 2000, contribuyendo a la evaluación sistemática de sistemas de acceso a la información, principalmente a través de experimentos en tareas competitivas.

Basándose en el formato introducido por primera vez en 2010, CLEF 2023 consta de una conferencia independiente revisada por pares que abarca una amplia gama de temas en los campos de la evaluación del acceso a

[1] https://clef2023.clef-initiative.eu/

la información multilingüe y multimodal, así como una serie de laboratorios y talleres diseñados para probar diferentes aspectos de los sistemas de recuperación de información monolingüe y multilingüe.

3.2. Tareas EXIST 2023

En la edición sobre la que se desarrolla este trabajo, EXIST 2023[2], se proponen tres tareas centradas en la clasificación de tuits tanto en inglés como en español: (i) identificación de sexismo, (ii) clasificación de la intención de la fuente y (iii) categorización de sexismo (Plaza et al., 2023).

3.2.1. Tarea 1 - Identificación de sexismo

La primera tarea consiste en una clasificación binaria en la que los sistemas deben determinar si dado un tuit, este expresa ideas relacionadas con el sexismo, tanto si es sexista en sí mismo como si describe una situación sexista en la que ocurre discriminación hacia las mujeres o si critica un comportamiento sexista. A continuación, se presentan ejemplos de mensajes sexistas y no sexistas (YES/NO).

Ejemplos de tuits con contenido sexista[3] (YES):

> *"@BestKabest Esta gringa sigue llorando por el gamergate, que "coincidencia" que tenga pronombres en su perfil"*

> *"you ever get so mad at your ex you send gamergate misogynist memes about about her to your 60 million fash followers https://t.co/TO1Br7kDJp"*

Ejemplos de tuits no sexistas (NO):

> *"Hay hembras de colibrí de cuello blanco (Florisuga mellivora) que adoptan una coloración similar al plumaje de los machos para evitar a los machos más agresivos. De esta forma, escapan de sus intentos de seducción y tienen más tiempo para alimentarse.https://t.co/MxbN4RyUMP"*

[2] http://nlp.uned.es/exist2023/

[3] Estos ejemplos se han extraído del conjunto de datos de entrenamiento donde las seis anotaciones tienen el mismo voto, YES o NO respectivamente.

"Anyway, as I said up this thread is for the handful of ppl that follow me that will care. For everyone else you can do/say whatever about me I really don't care anymore. Too many good ppl have been harassed or felt the need to leave bcuz of this fandom. I won't let them get to me"

3.2.2. Tarea 2 - Intención de la fuente

Esta tarea tiene como objetivo categorizar el mensaje según la intención del autor. En ella se propone una clasificación de los tuits en función de su intención al publicar el mensaje sexista. Esta distinción permite diferenciar el sexismo que está ocurriendo realmente en la red social, frente al sexismo que las mujeres están experimentando en otras situaciones, o el que se está denunciando en las redes sociales. Para ello se definen tres categorías que serán las que tengan que determinar los sistemas participantes:

- **Directo (direct)**: Mensaje sexista directo, donde la intención es escribir un mensaje sexista en sí mismo, como por ejemplo:

  "nunca vi a tanta gringa celebrando que los países pobres que nunca tenemos nada nos quedemos sin nada otra vez, pueden decir lo que quieran pero si tiene aires de microagresion y las odio https://t.co/xmlFevLzBN"

  "@laurenboebert Wow, you're like a less attractive Sarah Pailin - all ambition and NO BRAINS"

- **Reportado (reported)**: Mensaje sexista reportado, en el que la intención es informar y compartir una situación sexista experimentada por una mujer, o mujeres, en primera o tercera persona, como en:

  @DylanMc05 Sabe dónde vivo. No hay nada sexual, solo quiere joderme y hacerme daño. También físicamente. Le gusta abusar de mí y yo se lo he "permitido" demasiado tiempo... y no sé cómo pararle ahora.

  "no but this is me. i swore i was setting my daddy straight cause he use to piss me off with the "you gone have to cook for your husband." now look at me cant́ even cook for my damn self https://t.co/sVkVnAyGzd"

- **Juicio de valor (judgemental)**: Mensaje con juicio de valor. La intención es emitir un juicio, ya que el tuit describe situaciones o comportamientos sexistas con el objetivo de condenarlos.

  > *"espero que lo haya querido decir es que gracias a la pornografia y la cultura de la violacion, para muchas las relaciones sexuales pueden llegar a ser violentas por la idea de dominación/sumisión que esta tan erotizado que se use fuerza física contra nosotras https://t.co/6TXZpQiVxy"*

  > *"In other words; the games journo industry, who basically knew the entire time, are complicit? That can't be right. It's gotta be something to do with the misogynistic gamergate boogeymen. https://t.co/ql4UsT7UHI"*

3.2.3. Tarea 3 - Categorización de sexismo

El objetivo principal de la tarea es identificar automáticamente las diversas facetas de la vida de una mujer de las que son objeto los mensajes y comportamientos sexistas identificados en las redes sociales. Estas pueden abarcar los roles domésticos o de maternidad, oportunidades laborales, imagen sexual o las expectativas de vida, entre otros aspectos. Para ello, se definen cinco categorías que los sistemas deben tratar de detectar. Además, estas categorías no tienen por qué ser independientes, por lo que se trata de una tarea de multi-etiquetado, donde cada instancia puede ser etiquetada por una o más de las siguientes categorías:

- **Ideología-desigualdad (ideological-inequality)**: esta categoría engloba aquellos mensajes en redes sociales que desacreditan el movimiento feminista con la intención de menospreciar, ridiculizar y difamar la lucha de las mujeres en cualquier aspecto de sus vidas. Asimismo, abarca aquellos mensajes que rechazan la desigualdad entre hombres y mujeres, o que presentan a los hombres como víctimas de opresión basada en el género.

  > *"Pero noooo los jotos no son misoginos, pueden ser machos, gordofobicos, plumofobos, misoginos, pero también pueden no entender nada sobre la diferencia entre sexo/género/expresión de género https://t.co/b6xUNr1pvx"*

> *"Remember when and why the misogynists in positions of authority changed Women's studies to gender studies at universities all over the country?Remember what Feminists said would happen?I do. https://t.co/TySzlrbiBT"*

- **Estereotipado-dominación (stereotyping-dominance)**: donde la publicación expresa ideas falsas sobre la mujer en las que se sugiere que es más adecuada para cumplir determinados roles (ama de casa, esposa, cuidadora, amorosa, sumisa, etc.) o a la inversa, no adecuada para determinadas tareas (conducir, trabajo duro, etc.) e incluso donde se afirma que el hombre es superior a la mujer en algún sentido.

> *"Y recuerden que decir "No todos los hombres" no es más que una microagresión que demuestra tu necesidad de controlar la opinión de una mujer y someterla a los parámetros de lo que tú esperas que ella sea. https://t.co/O3bDl3pSRh"*

> *"@emrazz It's funny I saw someone create a thread of this but for men on what they would do with a day without women and the majority of responses were "enjoy the silence" and "play video games uninterrupted" but not one single comment about safety."*

- **Objetivificación (objectification)**: esta categoría abarca mensajes en los cuales las mujeres son presentadas como objetos, desvinculadas de su dignidad y aspectos personales. También se incluyen mensajes que asumen o describen ciertas cualidades físicas que las mujeres deben poseer para cumplir con los roles de género tradicionales. Por ejemplo, ideas que sugieren que las mujeres deben mantener un estándar e ideal de belleza, o críticas hacia la apariencia física de una mujer.

> *"... a las mujeres hay que tratarlas con delicadeza total... Y hasta donde ellas dejen.... Ahi esta el truco..."*

> *"Call me sexist but if I were a woman with big knockers this tweet would have garnished probably over 500 plus likes, so where is the reality?"*

- **Violencia sexual (sexual-violence)**: en esta categoría se engloban mensajes que involucran expresiones que pueden ser interpretadas como sugerencias, solicitudes o actos de acoso sexual, los cuales pueden abarcar desde comentarios explícitos hasta propuestas inapropiadas que incitan o promueven la violencia sexual. Esta categoría busca identificar y catalogar mensajes que presentan contenido ofensivo y peligroso relacionado con la violencia sexual, con el propósito de abordar y combatir este tipo de comportamientos inaceptables en las plataformas en línea.

  > *"Las denuncias por sumisión química suben un 36 % en Madrid. Nos están echando drogas en las copas, drogan a nuestras hijas, sobrinas y a nuestras amigas. Acabaremos con la cultura de la violación trabajando por el empoderamiento, la seguridad y la libertad de las mujeres. https://t.co/jXqwxZUaiV"*

  > *"People call me sexist but i aint any of that shit like bro, id totally fuck a woman compared to 6"3 pablo's ass that looks like its anal retraction could slice my cock like salami"*

- **Misoginia, violencia no sexual (misogyny-non-sexual-violence)**: en esta categoría se incluyen expresiones de odio y violencia dirigidas hacia las mujeres, abordando manifestaciones escritas que denotan un profundo desprecio o aversión hacia el género femenino. Estos mensajes pueden adoptar diversas formas, desde insultos y descalificaciones hasta amenazas y hostigamientos, y se caracterizan por perpetuar una actitud discriminatoria basada en el género.

  > *"No lleves esa falda, pareces una puta. Te falta calle, no sabes nada de la vida. Tus amigas son todas unas zorras, es mejor que no quedes con ellas. Tu hermana intenta ponerte contra mí, no sabe lo que dice. Cállate puta!!!!!!!"*

  > *"Don't be scared to spank women, they want to be spanked.Even those that don't like to be spanked, only think so because they haven't been spanked right.Spanking can Greatly increase the sensation, that + the feeling of being dominated by a strong Man= Jerking Orgasms"*

	Anotadores	Países	tuits
Entrenamiento	725	45	6.920
Desarrollo	113	23	1.038
Test	227	28	2.076

Tabla 3.1: Distribución del conjunto de datos EXIST 2023: En esta tabla se muestran para cada subconjunto de datos aportado por la competición, el número de anotadores que han participado, la diversidad geográfica y la cantidad de tuits que lo forman.

3.3. Conjunto de datos EXIST 2023

El conjunto de datos EXIST 2023, aunque de nueva creación, es similar en rasgos generales a los utilizados en ediciones anteriores. Está conformado por tuits tanto en inglés como en español, el conjunto de entrenamiento consta de más de 3,200 tuits por cada idioma, el conjunto de desarrollo incluye 500 tuits por cada idioma y el conjunto de pruebas contiene 1,000 tuits por cada idioma.

En esta edición, además, para garantizar perspectivas diversas y mitigar el sesgo en las etiquetas, cada tuit en el conjunto de datos es anotado por seis individuos reclutados a través del servicio Prolific[4]. El género de los anotadores (masculino/femenino[5]) y la edad (18-22 años, 23-45 años, +46 años) se tienen en cuenta durante el proceso de etiquetado. Como resultado, cada tuit está etiquetado por anotadores de género y grupos de edad diferentes.

Finalmente, el conjunto de datos abarca una amplia gama de tuits procedentes de diversos países e incluye anotaciones proporcionadas por un grupo diverso de anotadores. En la tabla 3.1 se muestra la distribución de los datos, destacando el número de anotadores, países y tuits para cada subconjunto (entrenamiento, desarrollo y test).

3.3.1. Formato del conjunto de datos

En la tabla 3.1 se observa que este conjunto de datos de la competición EXIST 2023 está distribuido en tres subconjuntos (entrenamiento, desarrollo y test). Estos son aportados por la organización en las diferentes fases del desarrollo de la competición.

El conjunto de datos, se distribuye en **formato JSON**, de tal manera que permite representar cada instancia como un objeto con los siguientes

[4] https://app.prolific.co

[5] Por razones de disponibilidad, solo se consideran los géneros masculino y femenino

atributos:

- **"id_EXIST"**: identificador único del tuit.
- **"lang"**: identificador del idioma del tuit ("en" o "es").
- **"tweet"**:texto del tuit.
- **"number_annotators"**: número de personas que han anotado el tuit.
- **"annotators"**: identificador único de cada uno de los anotadores de este tuit.
- **"gender_annotators"**: género de cada uno de los anotadores de este tuit.
- **"age_annotators"**: rango de edad al que pertenece cada uno de los anotadores de este tuit.
- **"labels_task1"**: conjunto de etiquetas que indican si el tuit de la instancia contiene expresiones sexistas o no ("YES" o "NO"). Hay una anotación por cada uno de los anotadores.
- **"labels_task2"**: conjunto de etiquetas (una por anotador) que indican la intención de quien publica el tuit. Solo disponible si el tuit ha sido etiquetado como "YES" en la tarea 1.
- **"labels_task3"**: conjuntos de etiquetas (una por anotador) que indican el o los tipos de sexismo encontrados en el tuit. Solo disponible si el tuit ha sido etiquetado como "YES" en la tarea 1.
- **"split"**: indicador del subconjunto del dataset al que pertenece el tuit (TRAIN, DEV, TEST).

De estos subconjuntos, los dos primeros (entrenamiento y desarrollo) son los que debe utilizar el participante para crear y entrenar sus sistemas. El subconjunto de test, por otro lado, únicamente tiene los campos **"id_EXIST"**, **"lang"**, **"tweet"**, y **"split"**, dado que será la labor de los sistemas desarrollados hallar los resultados para las tres tareas propuestas.

3.3.2. Formato de los outputs

Para participar en cada una de las tareas propuestas, el participante debe enviar hasta tres propuestas de resultados (outputs) basadas en el conjunto de **"test"**. El conjunto de resultados a presentar es un archivo en formato json en el que para cada una de las instancias, el sistema debe proveer de **hard labels** (etiquetas únicas) y/o **soft labels** (distribución de probabilidad). De este modo, los conjuntos de resultados para cada tarea deben ser:

- **Tarea 1: Identificación de sexismo.** Tal como se ha visto en la sección 3.2.1, esta se trata de una tarea de clasificación binaria, donde el sistema tiene que determinar si el tuit es sexista o no. Para ello, el conjunto de resultados debe tener, para cada instancia del dataset de **test**, un objeto JSON con los siguientes campos:

  - **"id_EXIST"**: identificador único del tuit.
  - **"hard_label"**: etiqueta única con los posibles valores "YES" o "NO".
  - **"soft_label"**: para cada una de las dos etiquetas posibles ("YES" y "NO"), se debe proporcionar una probabilidad de modo que su suma ha de ser 1.

  En la figura 3.1, se muestra un ejemplo de un objeto JSON con la estructura de un output de la tarea 1 con dos tuits.

- **Tarea 2: Intención de la fuente.** En la segunda tarea de la competición se debe determinar cuál es la intención de quien escribe el tuit, si es un mensaje sexista, si se trata de denunciar una situación sexista o si la intención es emitir un juicio de valor respecto a una situación sexista. Para ello, para cada instancia del subconjunto de test se debe proponer un resultado con los siguientes campos:

  - **"id_EXIST"**: identificador único del tuit.
  - **"hard_label"**: etiqueta única con los posibles valores "NO", "DIRECT", "REPORTED" o "JUDGEMENTAL".
  - **"soft_label"**: Para cada una de las cuatro etiquetas posibles ("NO", "DIRECT", "REPORTED" y "JUDGEMENTAL"), se

```
{
  "300002": {
    "hard_label": "YES",
    "soft_label": {
      "YES": 0.8333333333,
      "NO": 0.1666666667
    }
  },
  "300003": {
    "hard_label": "NO",
    "soft_label": {
      "NO": 1,
      "YES": 0
    }
  }
}
```


Figura 3.1: Ejemplo de un resultado para dos tuits de la tarea 1: en este ejemplo se muestran dos objetos *JSON* con los resultados *"hard_label"* donde se muestra la categoría predicha por el modelo y *"soft_label"* donde se encuentra la distribución de probabilidad de las dos etiquetas posibles (YES/NO) donde la suma de probabilidades ha de ser 1.0

debe proporcionar su probabilidad teniendo en cuenta que el sumatorio de probabilidades ha de ser 1.0.

En la figura 3.2, se muestra un ejemplo de un objeto JSON con la estructura de un fichero de resultados de la tarea 2 con dos tuits.

- **Tarea 3 - Categorización del sexismo.** En esta tarea, el sistema propuesto tiene la función de identificar la(s) etiqueta(s) correspondiente(s) entre las seis categorías posibles (NO, IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJETIFICATION, SEXUAL-VIOLENCE o MISOGYNY-NON-SEXUAL-VIOLENCE). A diferencia de las tareas previas, este escenario permite la asignación de más de una etiqueta por instancia. En consecuencia, la suma de probabilidades puede exceder de 1.0 en este contexto.

  - **"id_EXIST"**: identificador único del tuit.

```
{
  "300002": {
    "hard_label": "JUDGEMENTAL",
    "soft_label": {
      "JUDGEMENTAL": 0.5,
      "REPORTED": 0.3333333333,
      "NO": 0.1666666667,
      "DIRECT": 0
    }
  },
  "300003": {
    "hard_label": "NO",
    "soft_label": {
      "NO": 1,
      "DIRECT": 0,
      "REPORTED": 0,
      "JUDGEMENTAL": 0
    }
  }
}
```


Figura 3.2: Ejemplo de un resultado para dos tuits de la tarea 2: en este ejemplo se muestran dos objetos *JSON* con los resultados *"hard_label"* donde se muestra la categoría predicha por el modelo y *"soft_label"* donde se encuentra la distribución de probabilidad de las etiquetas posibles (DIRECT/REPORTED/JUDGEMENTAL/NO) donde la suma de probabilidades ha de ser 1.0.

- **"hard_label"**: lista de etiquetas con los posibles valores NO, IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJETIFICATION, SEXUAL-VIOLENCE o MISOGINY-NON-SEXUAL-VIOLENCE. Puede ser más de uno por cada instancia.
- **"soft_label"**: Para cada una de las seis etiquetas posibles (NO, IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJETIFICATION, SEXUAL-VIOLENCE o MISOGINY-NON-SEXUAL-VIOLENCE), se debe proporcionar una distribución de probabilidad teniendo en cuenta que el sumatorio de probabilidades puede ser superior a 1.0.

En la figura 3.3, se muestra un ejemplo de un objeto JSON con la estructura de un output de la tarea 3 con dos tuits.

```
{
  "300002": {
    "hard_label": [
      "IDEOLOGICAL-INEQUALITY",
      "STEREOTYPING-DOMINANCE",
      "MISOGYNY-NON-SEXUAL-VIOLENCE"
      ],
    "soft_label": {
      "IDEOLOGICAL-INEQUALITY": 0.3333333333,
      "STEREOTYPING-DOMINANCE": 0.3333333333,
      "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.5,
      "NO": 0.1666666667,
      "OBJECTIFICATION": 0.1666666667,
      "SEXUAL-VIOLENCE": 0
    }
  },
  "300004": {
    "hard_label": ["SEXUAL-VIOLENCE"],
    "soft_label": {
      "NO": 0.1666666667,
      "SEXUAL-VIOLENCE": 0.6666666667,
      "STEREOTYPING-DOMINANCE": 0.1666666667,
      "IDEOLOGICAL-INEQUALITY": 0,
      "MISOGYNY-NON-SEXUAL-VIOLENCE": 0,
      "OBJECTIFICATION": 0
    }
  }
}
```


Figura 3.3: Ejemplo de un resultado para dos tuits de la tarea 3: en este ejemplo se muestran dos objetos *JSON* con los resultados *"hard_label"* donde se muestra una lista de las categorías predichas por el modelo y *"soft_label"* donde se encuentra la distribución de probabilidad de las posibles categorias (NO/IDEOLOGICAL-INEQUALITY/STEREOTYPING-DOMINANCE/OBJETIFICATION/SEXUAL-VIOLENCE/MISOGINY-NON-SEXUAL-VIOLENCE). A diferencia de las dos tareas anteriores, en esta, al tratarse de una tarea de clasificación multi-etiqueta, la suma de probabilidades puede ser mayor de 1.0

3.4. Evaluación

Desde un punto de visto de métricas de evaluación, las tres tareas de la competición se pueden describir como:

- **Tarea 1 - Identificación de sexismo**: clasificación binaria mono-etiqueta.

- **Tarea 2 - Intención de la fuente**: clasificación jerárquica multiclase mono-etiqueta. La jerarquía de las clases tiene su primer nivel en si la publicación es o no sexista, y un segundo nivel para las tres clases exclusivas (directo/reportado/juicio de valor).

- **Tarea 3 - Categorización del sexismo**: clasificación jerárquica, multi-clase y multi-etiqueta. Del mismo modo que la segunda tarea, en esta el primer nivel se basa en si el contenido es o no sexista. En el segundo se determina la categoría del contenido sexista (ideología y desigualdad/dominación y estereotipado/objetivación/violencia sexual/misoginia y violencia no sexual). Además, las clases del segundo nivel no son mutuamente exclusivas, por lo que un tuit puede pertenecer a varias de ellas al mismo tiempo.

Para la evaluación de los sistemas en esta competición es necesario tener en cuenta que, al adoptarse en el paradigma *learning with disagreements*(Uma et al., 2021b), los conjuntos de datos de desarrollo, pruebas y validación están compuestos por múltiples anotaciones, concretamente seis, una por cada una de las cohortes o estratos definidos, y que bajo la perspectiva actual de herramientas de evaluación es necesario trabajar con formatos que permitan comparar la anotación del conjunto de datos respecto a la salida de los diferentes conjuntos de datos. Por tanto, es necesario considerar este paradigma desde los dos puntos de vista del proceso de evaluación:

(I) **Ground truth**: (verdad fundamental): en un escenario estricto *(hard)*, la variabilidad introducida por los anotadores se debe simplificar a un conjunto fijo de etiquetas que se asignan a cada instancia, por ejemplo con el uso de voto mayoritario. En cambio, en un escenario soft, el *gold standard* es el conjunto completo de las anotaciones, incluida su variabilidad. Por lo tanto, la métrica de evaluación debe tener en cuenta la proporción de anotadores que han elegido cada categoría. Las dos primeras tareas son problemas de una sola etiqueta, por lo que la suma de las probabilidades de cada clase debe ser uno. Pero en la tarea 3, que es de múltiples etiquetas, cada anotador puede seleccionar más de una categoría para un solo elemento. Por lo tanto, la suma de las probabilidades de cada clase puede ser mayor que uno.

(II) **System output**: (salida del sistema): en este caso, en un escenario estricto *(hard)*, los sistemas deben predecir una o más categorías para cada instancia. En un escenario "flexible" *(soft)*, el sistema debe predecir una probabilidad para cada categoría de las posibles en cada instancia. La puntuación de evaluación será máxima cuando las probabilidades predichas coincidan con las probabilidades reales en el

escenario flexible. Nuevamente, se ha de tener en cuenta que en la tarea 3, que es un problema de múltiples etiquetas, las probabilidades predichas por el sistema para cada una de las categorías no necesariamente suman uno.

Teniendo esto en cuenta, se consideran tres tipos de evaluación para cada una de las tareas:

- **Hard-hard**: el *gold standard* y la salida del sistema deben coincidir en el valor de las etiquetas. Para ello, en este enfoque se define un umbral para decidir las etiquetas finales en base a las diferentes etiquetas proporcionadas por los anotadores. En la tarea 1, se selecciona la clase elegida por más de tres anotadores; para la tarea 2, la clase elegida por más de dos anotadores; y en la tarea 3 (multi-etiqueta), se eligen las clases seleccionadas por más de un anotador. Los elementos sin una clase mayoritaria se excluyen de esta evaluación.

  La métrica oficial es ICM (Amigo y Delgado, 2022), aunque también se calcula F1. Para la tarea 1, se usa F1 para la clase positiva, mientras que en las tareas 2 y 3 se utiliza el promedio de F1 para todas las clases. Sin embargo, hay que tener en cuenta que F1 podría no ser la métrica ideal en este contexto, ya que no considera las relaciones entre las clases. Por ejemplo, un error entre no sexista y cualquiera de las subclases sexistas, y un error entre dos de las subclases positivas, se penalizan de manera similar, aunque el primero es un error más grave.

- **Hard-soft**: esta evaluación compara las categorías asignadas por el sistema con las probabilidades asignadas a cada categoría en el *gold standard*. La métrica oficial de evaluación en esta variante será ICM-soft (Plaza et al., 2023). Las probabilidades de las clases para cada instancia se calcularán en función de la distribución de etiquetas y el número de anotadores para esa instancia.

- **Soft-soft**: esta evaluación compara las probabilidades asignadas por el sistema con las probabilidades asignadas por el conjunto de anotadores humanos. Al igual que en el caso anterior, ICM-soft se utilizará como la métrica oficial de evaluación en esta variante.

Capítulo 4

Combinación de LLMs basados en Transformers con información socio-demográfica para detectar contenido sexista en redes sociales

En este capítulo se describe el sistema general propuesto para las tareas 1 y 2 de la competición EXIST 2023. También se describen las diferentes estrategias utilizadas tanto para la tarea de detección de sexismo y categorización de la intención de la fuente, como las utilizadas para tratar el desacuerdo de los anotadores.

4.1. Descripción general del sistema

Las tareas de la edición de 2023 de la competición EXIST se basan en el tratamiento de dos problemas principales, el primero es resolver la detección y clasificación del sexismo en redes sociales, y el segundo definir estrategias para tratar el desacuerdo inherente en el conjunto de datos aportado. En este caso, cada instancia del conjunto de datos tiene seis etiquetas basadas en las anotaciones proporcionadas por anotadores pertenecientes a seis co-

hortes distintas, tal como se explica en el apartado 3.3. Para tratar esos dos problemas, y en el marco de este trabajo, se propone un sistema basado en la combinación de diferentes estrategias que mejores resultados han obtenido en ediciones anteriores de esta competición en combinación con algunas de las técnicas observadas en el capítulo 2 sobre *Learning With Disagreement (LeWiDi)*.

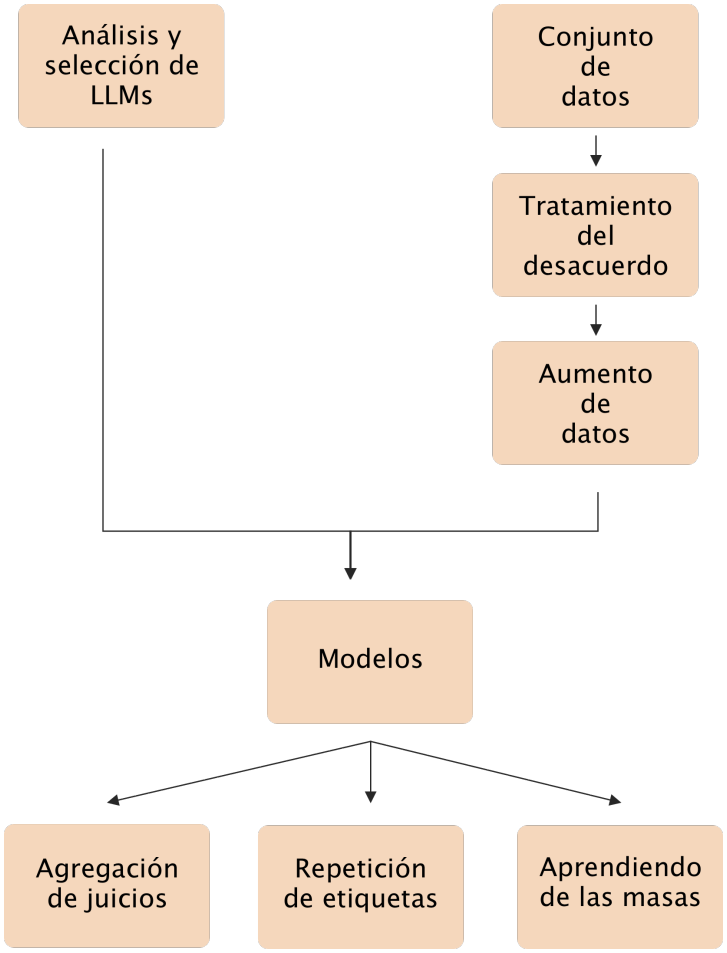

Figura 4.1: Esquema del sistema desarrollado para EXIST 2023: en este esquema se muestra como, de forma paralela, se hace una selección de modelos LLMs, por un lado, y se tratan el conjunto de datos aportado por la competición respecto al desacuerdo y aumento de datos por otro. La combinación de los modelos escogidos con las diferentes variantes del conjunto de datos definen finalmente tres sistemas que son los propuestos para la competición.

Tal como se puede observar en la figura 4.1, el sistema propuesto para este trabajo, se basa primero en el análisis y selección de grandes modelos del

lenguaje *"LLMs"* por sus siglas en inglés (*Large Language Models*) de los que se han seleccionado finalmente dos, uno para español y otro para inglés. De forma paralela, a partir del conjunto de datos aportado por la competición, se hace primero un tratamiento inicial del desacuerdo mediante diferentes estrategias de las definidas en el capítulo 2, obteniendo diferentes conjuntos de datos. A estos se les aplican a su vez diferentes estrategias de aumento de datos hasta obtener aproximadamente cuarenta variaciones del conjunto de datos original. Estas variaciones permiten ajustar los *LLMs* seleccionados previamente obteniendo diferentes modelos de detección de sexismo.

En función de la estrategia de *LeWiDi* en que se basen, algunos de estos modelos resultantes son de detección directa de sexismo. Otros, en cambio, se centran en una de las cohortes definidas, por lo que a los resultados obtenidos en este caso, se añade un procesamiento de agregación de los datos con diferentes configuraciones en función de los rasgos de género y edad de las propias cohortes.

4.2. Tecnología Transformers

A partir de los resultados de ediciones anteriores de la competición EXIST que se muestran en el capítulo 2, se puede observar que la mayor parte de los sistemas con mejores resultados están basados de una u otra manera en la tecnología Transformer. Esta tecnología es un enfoque revolucionario en el campo del procesamiento del lenguaje natural (PLN) y tiene un gran impacto en una amplia variedad de aplicaciones de inteligencia artificial. Fue introducida por primera vez en el artículo (Vaswani et al., 2017) y desde entonces ha evolucionado para ser la base de muchos modelos de vanguardia en PLN.

El impacto de esta tecnología en el ámbito del procesamiento del lenguaje, se puede apreciar en gran medida observando los avances en esta área durante los últimos 10 años. En la imagen 4.2 se observa cómo los diferentes avances en el área de Inteligencia Artificial aplicados al Procesamiento del Lenguaje Natural crean diferentes disrupciones en el estado del arte. En 2014 los equipos de Bengio (Bahdanau, Cho, y Bengio, 2014) y Google (Sutskever, Vinyals, y Le, 2014) propusieron la arquitectura Secuencia a Secuencia (Seq2Seq), que permite longitudes variables en las secuencias de entrada y salida de un sistema. Esta nueva arquitectura supuso un avance en el estado

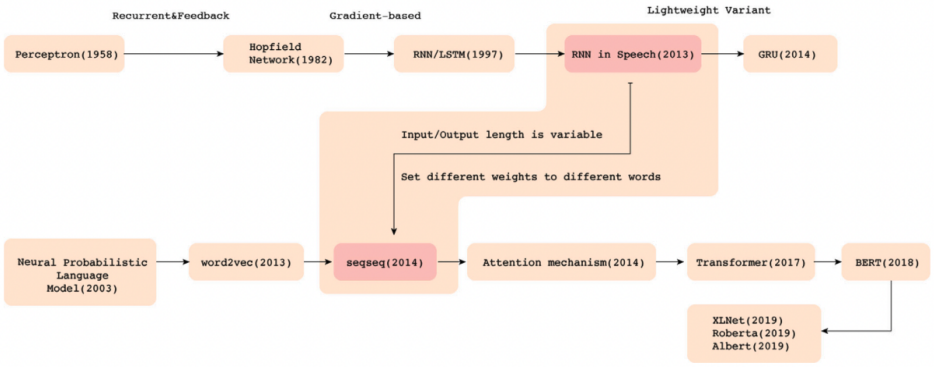

Figura 4.2: Historia del desarrollo del modelo secuencial (Shao et al., 2022).

del arte, siendo útil en traducción, chatbots, análisis de sintaxis, resumen de texto, entre otros.

Sin embargo, el enfoque Seq2Seq tenía sus deficiencias, como el hecho de comprimir toda la información de una secuencia de preguntas en un vector pequeño a través del codificador dificulta la descodificación sin perder información. Por lo que, casi al mismo tiempo, el equipo de Bengio propuso el mecanismo de atención (Bahdanau, Cho, y Bengio, 2014). Este mecanismo asigna diferentes pesos a diferentes palabras, lo que produce una secuencia de vectores de diferentes longitudes en lugar de una semántica intermedia de longitud fija. Esta mejora permite una utilización más completa de la información de la secuencia de entrada en cada paso de generación, resolviendo el problema de pérdida de información.

Posteriormente, Google propuso el marco Transformer en el popular artículo "Attention is all you need" (Vaswani et al., 2017), basado en el modelo Seq2Seq, para abordar sus defectos, dado que la principal limitación de Seq2Seq era la compresión de toda la información del lado del codificador en un vector de longitud fija, lo que resultaba en pérdida de información. Transformer mejoró esto al introducir la atención *multiHead* y un módulo de auto-atención que permite que las secuencias fuente y objetivo se "autoasocien", lo cual enriquece la representación de la secuencia fuente y objetivo. Además, Transformer mejoró la capacidad de cómputo paralelo, superando en gran medida a los modelos de la serie Seq2Seq (Shao et al., 2022).

4.3. Selección de modelos preentrenados

En las ediciones anteriores de EXIST, sobretodo en la edición de 2021, se observa una remarcable diferencia entre los sistemas basados en la tecnología transformers frente a aquellos que se basan en técnicas tradicionales de aprendizaje automático, como SVM (Support Vector Machine) y RF (Random Forest) tal como se explica en la sección 2.4.1. Teniendo esto en cuenta, se ha optado por utilizar modelos pre-entrenados basados en esta tecnología para el planteamiento de las diferentes aproximaciones que se han valorado durante el desarrollo de este trabajo.

Para establecer una línea de base sobre los múltiples modelos que se pueden encontrar en p.e. HuggingFace[1], primero se han seleccionado y examinado los modelos que se han utilizado en ediciones anteriores de la competición, para a continuación, entrenarlos y evaluarlos en la tarea 1 de EXIST 2021 de clasificación binaria para cada idioma.

Estos modelos son:

- **xlm-roberta-base** (Liu et al., 2019a): XLM-RoBERTa es una versión multilingüe de RoBERTa. Está preentrenado en 2.5TB de datos filtrados de CommonCrawl que contienen 100 idiomas. Este modelo está implementado con 12 capas y 270 millones de parámetros.

- **xlm-roberta-large** (Liu et al., 2019a) es otra versión multilingüe de RoBERTa. Está pre-entrenado en 2.5TB de datos filtrados de CommonCrawl que contienen 100 idiomas. Este modelo está implementado con 24 capas y 355 Millones de parámetros.

- **bert-base-multilingual-cased** (Devlin et al., 2018b): está basado en las 104 lenguas principales utilizando Wikipedia como fuente de datos y se enfoca en el Modelado de Lenguaje Enmascarado (MLM). Fue presentado en la publicación (Devlin et al., 2018c) y lanzado inicialmente en un repositorio público en GitHub[2]. Cabe destacar que este modelo es sensible a las mayúsculas y minúsculas, lo que significa que distingue entre "english" y "English".

- **distilbert-base-multilingual-cased** (Victor Sanh, 2023a): Este es un modelo de transformers más pequeño y rápido que BERT, el cual

[1] https://huggingface.com
[2] https://github.com/google-research/bert

fue pre-entrenado en el mismo corpus de manera auto-supervisada, utilizando el modelo base de BERT como maestro o "teacher". Esto significa que fue pre-entrenado solo en textos crudos, sin etiquetado humano de ninguna manera (razón por la cual puede utilizar una gran cantidad de datos públicamente disponibles), mediante un proceso automático para generar entradas y etiquetas a partir de esos textos utilizando el modelo base de BERT.

- **roberta-base-bne** (Fandiño et al., 2022a; Liu et al., 2019b): es un modelo de lenguaje basado en Transformer para el idioma español que se utiliza en tareas donde se enmascara parte del texto. Está basado en el modelo base RoBERTa y ha sido preentrenado utilizando el corpus en español más grande conocido hasta la fecha, con un total de 570 GB de texto limpio y sin duplicados procesados para este trabajo, recopilado de rastreos web realizados por la Biblioteca Nacional de España entre 2009 y 2019.

- **roberta-large-bne** (Fandiño et al., 2022b) está basado en XLM-RoBERTa-large para el idioma español, preentrenado utilizando el corpus español más grande conocido hasta la fecha, con un total de 570 GB de texto limpio y duplicado procesado para este trabajo, recopilado a partir de rastreos web realizados por la Biblioteca Nacional de España desde 2009 hasta 2019.

- **bertin-roberta-base-spanish** (la Rosa et al., 2022): BERTIN es una serie de modelos para el idioma español basados en BERT. Este proyecto se enfocó en pre-entrenar un modelo RoBERTa-base desde cero durante el evento de la Comunidad Flax/JAX, utilizando recursos gratuitos de Google Cloud, específicamente TPUv3-8. El objetivo era aprovechar las implementaciones de Flax de la librería de Huggingface para realizar el entrenamiento del modelo.

- **bert-base-spanish-wwm-cased** (Cañete et al., 2020b): es un modelo basado en BERT entrenado sobre un gran corpus en español con la técnica Whole Word Masking, obteniendo un tamaño similar a BERT-Base.

- **distillbert-base-spanish-uncased** (Victor Sanh, 2023b): se trata del mismo modelo distilbert-base-uncased entrenado con The Larg Spa-

nish Corpus (Cañete, 2019), que es una recopilación de 15 corpus españoles sin etiquetar que abarcan desde Wikipedia hasta notas del parlamento europeo.

Una vez realizada la experimentación con todos estos modelos sobre los conjuntos de datos de EXIST 2021, se obtienen dos comparativas: la primera, a partir de los resultados para los tuits en español, se presentan en la tabla 4.1; mientras que la Tabla 4.2 muestra los resultados para los tuits en inglés.

Model	**F1**
xlm-roberta-base	0.703
xlm-roberta-large	0.447
bert-base-multilingual-cased	0.701
distilbert-base-multilingual-cased	0.706
PlanTL-GOB-ES/roberta-base-bne	**0.751**
PlanTL-GOB-ES/roberta-large-bne	0.741
bertin-project/bertin-roberta-base-spanish	0.747
dccuchile/bert-base-spanish-wwm-cased	0.709
CenIA/distillbert-base-spanish-uncased	0.713

Tabla 4.1: Revisión de LLMs en español: Esta tabla muestra los resultados de evaluar cada uno de los modelos seleccionados frente a la tarea 1 de EXIST 2021

Model	**F1**
xlm-roberta-base	0.652
xlm-roberta-large	0.676
bert-base-multilingual-cased	0.721
distilbert-base-multilingual-cased	0.701
roberta-large	0.288
distilbert-base-uncased	**0.741**
bert-base-uncased	0.733

Tabla 4.2: Revisión de modelos en inglés: Esta tabla muestra los resultados de evaluar cada uno de los modelos seleccionados frente a la tarea 1 de EXIST 2021

En las tablas 4.1 y 4.2, se puede observar que los mejores resultados se obtienen con el modelo **PlanTL-GOB-ES/roberta-base-bne** para los contenidos en español y **distilbert-base-uncased** para los contenidos en inglés. Es por ello que se seleccionan estos modelos como base en los próximos experimentos.

4.4. Tratamiento del desacuerdo: aproximación inicial

En el capítulo 3 se hace referencia a que los conjuntos de datos aportados para la competición tienen un formato sin *gold standard*; es decir, una única etiqueta para cada tarea. En cada instancia se aporta como etiqueta seis anotaciones realizadas cada una por un anotador, representante de su cohorte.


```
{
  "100030": {
    "id_EXIST": "100030",
    "lang": "es",
    "tweet": "@Quora Is a great example of this, they don't care if I have being
      supplanted, harassed or else, even with evidence the problem was me. Por
      supuesto en Quora español. Hay 1 tipo con cientos de cuentas ofensivas y
      sigue ahí. Y el clonador de cuentas... https://t.co/cRNZ0VNnUn https://t.co
      /oQnHzyOrsN",
    "number_annotators": 6,
    "annotators": [ "Annotator_109","Annotator_110","Annotator_111"
      ,"Annotator_112","Annotator_113","Annotator_114"],
    "gender_annotators": ["F","F","F","M","M","M"],
    "age_annotators": ["18-22","23-45","46+","46+","23-45","18-22"],
    "labels_task1": ["YES","YES","NO","YES","NO","NO"],
    "labels_task2": ["REPORTED","REPORTED","-","REPORTED","-","-"],
    "labels_task3": [
      ["STEREOTYPING-DOMINANCE","OBJECTIFICATION","SEXUAL-VIOLENCE","MISOGYNY-NON
        -SEXUAL-VIOLENCE"],
      ["SEXUAL-VIOLENCE"],
      ["-"],
      ["OBJECTIFICATION"],
      ["-"],
      ["-"]
    ],
    "split": "TRAIN_ES"
  }
}
```


Figura 4.3: Ejemplo de anotación de un tuit: en este ejemplo, se muestra la estructura de una instancia dentro del conjunto de datos aportado por la competición. En cada instancia, se aporta el identificador de la misma, el idioma, el texto del tuit, el número de anotadores, el género y rango de edad de cada uno y las anotaciones realizadas por cada uno de ellos. En la tarea 3 se puede observar que un anotador puede anotar el tuit en más de una categoría.

Es por ello que previamente a cualquier otra labor se ha tenido que realizar un preprocesamiento del dataset con el fin de trata el desacuerdo entre anotadores. Para ello, se han utilizado dos de las diferentes estrategias de las descritas en la subsección de *Learning With Disagreement* 2.5.2 del capítulo de estado del arte.

- **Transformación por agregación de juicios:** la primera estrategia se basa principalmente en la **agregación de juicios** tal y como se presenta en el trabajo de (Dawid y Skene, 1979), definiendo una etiqueta única de cada instancia basada en voto mayoritario pero conservando la distribución de probabilidad de las etiquetas.

- **Repetición de etiquetas**: teniendo en cuenta que el conjunto de datos está etiquetado por seis categorías de usuarios definidas por género (masculino y femenino) y edad (18-22 años, 23-45 años, +46 años), tal como se muestran en la tabla 2.3 se emplea la técnica de repetición de etiquetas ("repeated labeling") descrita anteriormente. Para ello, se procede a replicar cada instancia por cada cohorte, asignando a cada nueva instancia la anotación de cada una de las cohortes. De este modo se obtiene un conjunto de datos con una etiqueta única para cada instancia, pero el texto y metadatos de la instancia replicado, consiguiendo así transferir el desacuerdo de los anotadores directamente al modelo. En la tabla 4.3 se puede ver un ejemplo de la primera instancia del conjunto de datos de entrenamiento que ha sido tratada con esta técnica.

id_EXIST	lang	tweet	annotators	gender_annotators	age_annotators	labels_task1	labels_task2
100001	es	@TheChiflis Ignora al otro, es un capullo.El p...	Annotator_1	F	18-22	YES	REPORTED
100001	es	@TheChiflis Ignora al otro, es un capullo.El p...	Annotator_2	F	23-45	YES	JUDGEMENTAL
100001	es	@TheChiflis Ignora al otro, es un capullo.El p...	Annotator_3	F	46+	NO	-
100001	es	@TheChiflis Ignora al otro, es un capullo.El p...	Annotator_4	M	18-22	YES	REPORTED
100001	es	@TheChiflis Ignora al otro, es un capullo.El p...	Annotator_5	M	23-45	YES	JUDGEMENTAL
100001	es	@TheChiflis Ignora al otro, es un capullo.El p...	Annotator_6	M	46+	YES	REPORTED

Tabla 4.3: Ejemplo de instancia tratada con "repeated labeling": en esta tabla se puede observar que se trata de una única instancia en el conjunto de datos original que ha sido replicada seis veces, una por cada una de las cohortes que se han definido a partir de los datos socio-demográficos de los anotadores (género y edad).

4.5. Aumento de datos (data augmentation)

Independientemente de la estrategia empleada para tratar el desacuerdo en los conjuntos de datos, el tamaño de los conjuntos de datos aportados (6,920 instancias para el conjunto de entrenamiento y 2,076 para el de desarrollo) puede no ser suficiente para obtener resultados óptimos. Es por ello que se han contemplado diferentes opciones de aumento de datos para crear conjuntos de datos de entrenamiento de mayor tamaño.

Por tanto, en el segundo paso del desarrollo del sistema propuesto se han planteado diferentes estrategias para el aumento de datos a partir de los conjuntos de datos iniciales de entrenamiento y desarrollo:

- **Combinación con EXIST 2021**: al conjunto de datos obtenido mediante agregación de juicios se le ha añadido el conjunto de datos de EXIST 2021, adaptándolo al mismo formato.

- **Aumento mediante traducción de EXIST 2023** la siguiente variación se basa en una de las más utilizadas entre las propuestas presentadas en las ediciones de EXIST 2021 y EXIST 2022, y es el aumento de datos mediante traducción. Para ello, se han utilizado los modelos **'Helsinki-NLP/opus-mt-es-en'**[3] para traducir los tuits de español a inglés y **'Helsinki-NLP/opus-mt-en-es'**[4] para traducir los tuits de inglés a español. En este caso, también se ha mantenido inicialmente la perspectiva de unificar por voto mayoritario las diferentes anotaciones de cada instancia y los resultados obtenidos al evaluarlo demuestran que los modelos desarrollados con este nuevo dataset ampliado son mejores que los obtenidos con el conjunto de datos original.

A partir de la combinación de las estrategias de gestión inicial del desacuerdo y las de aumento de datos, se generan una serie de conjuntos de datos que serán los utilizados para ajustar los modelos preentrenados basados en la tecnología transformer seleccionados previamente. En la tabla 4.4, se muestran las diferentes combinaciones desarrolladas y el número de instancias resultantes obtenidas para cada uno.

[3] https://huggingface.co/Helsinki-NLP/opus-mt-es-en
[4] https://huggingface.co/Helsinki-NLP/opus-mt-en-es

Gestión desacuerdo	Idiomas	Aumento de datos	Instancias
Agregación de juicios	ES	Dataset EXIST 2023 inicial	3.660
		EXIST 2021[5]	7.201
		Traducción	7.320
		EXIST 2021 + traducción	14.402
	EN	Dataset EXIST 2023 inicial	3.660
		EXIST 2021	3.541
		Traducción	7.320
		EXIST 2021 + Traducción	14.402
	ES + EN	Dataset EXIST 2023 inicial	6.920
		EXIST 2021	13.897
		Traducción	13.840
		EXIST 2021 + Traducción	27.794
Repeated Labeling	ES	Dataset EXIST 2023 inicial	21.960
		Traducción	43.920
	EN	Dataset EXIST 2023 inicial	21.960
		Traducción	43.920
	ES+EN	Dataset EXIST 2023 inicial	41.520
		Traducción	83.040

Tabla 4.4: Variaciones de conjuntos de datos: en esta tabla se recogen las diferentes variaciones creadas del conjunto de datos original a partir de la combinación de los diferentes tratamientos hechos del desacuerdo entre anotadores junto con las diferentes estrategias de aumento de datos y en función del idioma.

4.6. Creación de modelos

En esta sección se describe la creación de diferentes modelos basados en los modelos preentrenados descritos en la sección 4.3 a los que se les aplica *finetuning* con los conjuntos de datos descritos en el apartado anterior.

Finetuning en el contexto de modelos Transformers se refiere a ajustar un modelo preentrenado en tareas específicas o dominios para mejorar su rendimiento en una tarea particular. Los modelos Transformer son conocidos por su capacidad para aprender patrones complejos en datos secuenciales y han demostrado ser eficaces en una variedad de tareas de procesamiento de lenguaje natural (NLP), como traducción automática, resumen de texto, y clasificación de texto, entre otras.

El proceso de "fine-tuning" generalmente implica tomar un modelo Transformer preentrenado en un conjunto de datos grande y ajustarlo a un conjunto de datos más pequeño y específico para la tarea que se desea abordar. Esto se hace mediante el entrenamiento adicional del modelo en la tarea específica con un conjunto de datos de entrenamiento más pequeño y específico.

El beneficio de utilizar este proceso de "fine-tuning" en modelos Transformer es que pueden aprovechar el conocimiento aprendido durante el preentrenamiento en datos masivos, lo que permite una adaptación eficiente a tareas específicas con menos datos de entrenamiento. Esto es particularmente útil en escenarios donde no hay suficientes datos para entrenar un modelo desde cero para la tarea deseada.

En las tablas 4.1 y 4.2 del apartado 4.3 se muestran los modelos que se han examinado para el desarrollo de este trabajo, entre todos ellos, se seleccionan los modelos con mejores resultados obtenidos: **'PlanTL-GOB-ES/roberta-base-bne'** para los contenidos en español y **distilbert-base-uncased** para aquellos en inglés.

Por otro lado, en la tabla 4.4 se muestra un listado de los conjuntos de datos generados a partir del original de la competición, inclusive, teniendo en cuenta la gestión del desacuerdo incluido y estratégias de aumento de datos.

A partir de los conjuntos de datos definidos en la tabla 4.4 y los modelos preentrenados seleccionados, se plantean tres estrategias principales con las que abordar las tareas, la primera basada en los conjuntos donde el desacuerdo se ha convertido en una única etiqueta mediante **agregación**, la segunda para los conjuntos tratados con **"repeated labeling"** y finalmente la tercera denominada **"Learning from crowds"**

4.6.1. Agregación de juicios (judgement aggregation)

El sistema basado en agregación de juicios, se implementa a partir de los conjuntos de datos en los que el tratamiento del desacuerdo se ha basado en la agregación por voto mayoritario, tal como se describe en la sección 4.4. De este modo y tal y como se explica en (Dawid y Skene, 1979), la etiqueta agregada se define por aquella que tiene mayor número de votos. Al ser seis el número de anotaciones por cada instancia, se decide que en caso de que el número de votos positivos y negativos sea equivalente, establecer el valor "NO" por defecto.

De esta manera se consigue un conjunto de datos principal en el que se ha eliminado el desacuerdo y se obtiene una etiqueta única para cada instancia, de modo que se pueden utilizar para afinar los modelos preentrenados escogidos previamente.

En la tabla 4.4 se puede observar que los conjuntos de datos creados por

Idiomas	Aumento de datos	Modelo preentrenado	ID
ES	-	Roberta	ag-1
	EXIST 2021	Roberta	ag-2
	Traducción	Roberta	ag-3
	EXIST 2021 + traducción	Roberta	ag-4
EN	-	Distilbert	ag-5
	EXIST 2021	Distilbert	ag-6
	Traducción	Distilbert	ag-7
	EXIST 2021 + Traducción	Distilbert	ag-8
ES + EN	-	Roberta	ag-9
		Distilbert	ag-10
	EXIST 2021	Roberta	ag-11
		Distilbert	ag-12
	Traducción	Roberta	ag-13
		Distilbert	ag-14
	EXIST 2021 + Traducción	Roberta	ag-15
		Distilbert	ag-16

Tabla 4.5: Modelos entrenados con agregación de juicios

agregación de juicios están distribuidos por idioma (español, inglés o ambos) y por método de aumento de datos (sin aumento, utilizando el conjunto de datos de la edición de 2021 de EXIST o una combinación de las dos).

La combinación de conjuntos en dos lenguajes, con las diferentes estrategias de aumento de datos y el entrenamiento con modelos pre-entrenados genera finalmente 16 modelos basados en agregación por voto mayoritario. Todos ellos son evaluados con el conjunto de datos de desarrollo aportado por la organización de la competición y se extrae que, para esta aproximación, el modelo **ag-14** basado en conjunto de datos multilingüe con aumento de datos por traducción y entrenado con el modelo **Distilbert** es el que obtiene mejores resultados.

En la tabla 4.5 se muestran los modelos creados a partir de las diferentes combinaciones de idioma, estrategia de aumento de datos y modelo preentrenado.

4.6.2. Repetición de etiquetado (repeated labeling)

En el capítulo 3 de este trabajo, se explica que los conjuntos de datos aportados para la competición EXIST 2023 están conformados por instancias anotadas por seis anotadores de seis cohortes diferentes en función de su edad y género. La edad, agrupada por los conjuntos 18-22, 23-45, 46 o más, y el género, distinguiendo entre hombres y mujeres.

Otra de las propuestas que se hacen en el capítulo 2 para gestionar el

Idiomas	Aumento de datos	Modelo preentrenado	ID
ES	-	Roberta	rl-1
	Traducción	Roberta	rl-2
EN	-	Distilbert	rl-3
	Traducción	Distilbert	rl-4
ES+EN	-	Roberta	rl-5
		Distilbert	rl-6
	Traducción	Roberta	rl-7
		Distilbert	rl-8

Tabla 4.6: Listado modelos creados por repeated labeling

desacuerdo entre anotadores se trata de **repeated labeling** que pertenece al conjunto de estrategias basadas en la idea de que no tiene porqué existir una única etiqueta para cada caso, sino que se trata de capturar el conocimiento global de los anotadores a partir de la distribución de etiquetas seleccionadas por cada uno de ellos.

En este planteamiento, se utilizan los conjuntos de datos definidos en la tabla 4.4 basados en **repeated labeling** para entrenar directamente los dos modelos, obteniendo de esta forma modelos en los que el desacuerdo entre anotadores está incluido.

La tabla 4.6 muestra el listado de modelos planteados a partir de los conjuntos de datos basados en "repeated labeling". Por tiempo y capacidad de cómputo, de los modelos definidos en la tabla se hace una primera evaluación que determina que los modelos monolingües aumentados por traducción dan mejores resultados, por lo que son los seleccionados para ser utilizados tanto en esta estratégia como en la de **Learning from crowds** (ver sección 4.6.3).

4.6.3. Aprendiendo de las masas (Learning from crowds)

Otra de las estrategias que se ha tenido en cuenta en la definición del sistema para este trabajo es la de entrenar un clasificador a partir del desacuerdo de los anotadores. De las diferentes opciones que se encuentran en la bibliografía, para este trabajo se ha decidido realizar una aproximación similar a la definida por (Rodrigues y Pereira, 2018), creando un modelo para cada anotador con la probabilidad de cada etiqueta y agregando estos resultados en diferentes configuraciones. Teniendo en cuenta que pese a que cada instancia es anotada por seis anotadores diferentes, el conjunto de da-

tos general (incluyendo los 3 splits train/dev/test) se ha anotado con hasta 1.064 anotadores de todo el mundo, se contempla que no hay suficientes anotaciones por cada anotador como para utilizar este planteamiento. Por tanto, se define una aproximación basada en las propias cohortes, es decir, entrenar un modelo para cada cohorte y como el resultado tiene que ser único, plantear diferentes opciones de agregación. En la tabla 4.7 se muestra el listado de los modelos creados para cada cohorte.

Idioma	Edad	Género	Modelo	ID
ES	18-22	M	Roberta	coh-1
		F	Roberta	coh-2
	23-45	M	Roberta	coh-3
		F	Roberta	coh-4
	46+	M	Roberta	coh-5
		F	Roberta	coh-6
EN	18-22	M	Distilbert	coh-7
		F	Distilbert	coh-8
	23-45	M	Distilbert	coh-9
		F	Distilbert	coh-10
	46+	M	Distilbert	coh-11
		F	Distilbert	coh-12

Tabla 4.7: Modelos basados en cohortes

Con la creación de estos 12 modelos, 6 para inglés y 6 para español, se realiza una predicción para cada instancia utilizando los 6 del idioma de la misma instancia, obteniendo como resultado una predicción por cada cohorte para el mismo tuit. En la figura 4.4 se muestra el resultado para la primera instancia (100001) del conjunto de datos de entrenamiento para la tarea 1 de clasificación binaria de la competición.

Une vez obtenidos los resultados para todas las instancias del conjunto de datos que se pretende predecir, dado que la competición exige que el resultado del sistema sea por un lado una única etiqueta ("hard_label") y por otro un único valor de la probabilidad de dicha etiqueta ("soft_label") es preciso agregar estos resultados en uno único con ambos valores. Para ello se plantean diferentes estrategias de agregación:

- Etiquetado por **voto mayoritario**: este método calcula la etiqueta que ha sido escogida por la mayoría de los modelos de las diferentes cohortes. En caso de empate, se ha resuelto utilizando la etiqueta **"NO"** por defecto para la primera tarea y **"JUDGEMENTAL" -**

```
{
  "F_18_22": [
    { "label": "NO", "score": 0.9990222454071045 },
    { "label": "YES", "score": 0.0009777863742783666 }
  ],
  "F_23_45": [
    { "label": "YES", "score": 0.9706707000732422 },
    { "label": "NO", "score": 0.02932928502559662 }
  ],
  "F_46": [
    { "label": "NO", "score": 0.9904109239578247 },
    { "label": "YES", "score": 0.009589116089046001 }
  ],
  "M_18_22": [
    { "label": "NO", "score": 0.9985015392303467 },
    { "label": "YES", "score": 0.0014985214220359921 }
  ],
  "M_23_45": [
    { "label": "NO", "score": 0.9767223000526428 },
    { "label": "YES", "score": 0.02327771857380867}
  ],
  "M_46": [
    {"label": "NO", "score": 0.9986475110054016},
    {"label": "YES", "score": 0.0013525356771424413}
  ]
}
```


Figura 4.4: Ejemplo del resultado obtenido por los seis modelos basados en cohortes: en este ejemplo se muestran seis objetos *JSON* correspondientes a las seis cohortes en las que se ha basado la anotación del conjunto de datos de EXIST 2023. La primera letra de los índices hace referencia al género (Masculino/Femenino) y los números a los rangos de edad (18 a 22, 23 a 45 y mayores de 46 años).

"REPORTED" - "DIRECT" como orden de preferencia para la segunda.

- Etiquetado por **voto basado en género**: este método calcula para cada instancia, el voto mayoritario para cada género y devuelve aquel que tenga más votos en caso de que no coincidan. Para los casos de empate se han planteado las dos variantes donde se establece el voto de los modelos basados en la cohorte de las mujeres, por un lado, y la cohorte de los hombres, por otro.

- Etiquetado por **voto basado en edad**: este método calcula para cada instancia, el número de modelos que han inferido cada etiqueta y compara los resultados. Dado que hay tres grupos de edad, no se contempla el caso de empate.

- Distribución de **probabilidad media**: este método calcula la probabilidad media para cada etiqueta obtenida a partir de los seis modelos,

y devuelve la distribución de probabilidad media para cada etiqueta.

- Distribución de **probabilidad por género**: este método calcula la etiqueta más probable para cada género, sumando las puntuaciones de cada modelo según el género de la cohorte en que esté basado. Se selecciona la etiqueta con la probabilidad más alta. Dado que se basa en la probabilidad, se asume que es muy difícil que se presente una situación de empate.

- Distribución de **probabilidad por edad**: este método calcula la etiqueta más probable para cada rango de edad sumando las puntuaciones del modelo de cada rango. Se selecciona la etiqueta con mayor probabilidad.

- Distribución de **probabilidad por cohorte**: este método calcula la probabilidad de cada etiqueta/cohorte considerando la probabilidad generada por el modelo. Luego se selecciona la etiqueta con la mayor probabilidad y se devuelve como predicción final.

Finalmente, de la combinación de los modelos preentrenados seleccionados, los métodos de agregación descritos y la configuración de estos mismos en caso de empate, resulta en una combinación de experimentación que se resume en la tabla 4.8

Aumento de datos	Método agregación	Estrategia en empate	ID sistema
NO	Voto mayoritario	Sin empate	sys-lfc-1
	Voto basado en género	Preferencia hombres	sys-lfc-2
	Voto basado en género	Preferencia mujeres	sys-lfc-3
	Voto basado en edad	Sin empate	sys-lfc-4
	Probabilidad media	Sin empate	sys-lfc-5
	Probabilidad por género	Sin empate	sys-lfc-6
	Probabilidad por edad	Sin empate	sys-lfc-7
	Probabilidad por cohorte	Sin empate	sys-lfc-8
Traducción	Voto mayoritario	Sin empate	sys-lfc-9
	Voto basado en género	Preferencia hombres	sys-lfc-10
	Voto basado en género	Preferencia mujeres	sys-lfc-11
	Voto basado en edad	Sin empate	sys-lfc-12
	Probabilidad media	Sin empate	sys-lfc-13
	Probabilidad por género	Sin empate	sys-lfc-14
	Probabilidad por edad	Sin empate	sys-lfc-15
	Probabilidad por cohorte	Sin empate	sys-lfc-16

Tabla 4.8: Listado de variaciones de sistemas basados en *Learning from crowds*

Capítulo 5

Evaluación y discusión

En el ámbito de clasificación de contenidos mediante modelos de lenguaje, la evaluación de estos sistemas adopta un papel crucial para determinar la eficacia y robustez de las soluciones propuestas. Más aún destacar esta necesidad en las tareas relacionadas con detección de contenido tóxico en cualquiera de sus formas, dado los perjuicios que puede producir un sistema que, por una evaluación deficiente, genere errores de interpretación en entornos en producción con millones de usuarios activos.

En este capítulo se presentan tanto la metodología de evaluación de los sistemas definidos en el capítulo 4, como los resultados obtenidos tanto en las evaluaciones previas con, el conjunto de datos de desarrollo, como en la evaluación final de la primera y segunda tareas de detección y clasificación de sexismo en la competición EXIST 2023. Así mismo, se interpretarán y discutirán los resultados obtenidos, teniendo en cuenta diferentes configuraciones y las implicaciones de los hallazgos en el contexto de desacuerdo entre anotadores.

5.1. Metodología de evaluación

El planteamiento de este trabajo consiste en abordar las dos primeras tareas de la competición EXIST 2023, las cuales desde un punto de vista de evaluación se describen como una tarea de clasificación binaria monoetiqueta en el caso de la primera tarea, y una tarea de clasificación jerárquica multi-clase para la segunda.

Además, se ha de tener en cuenta para la evaluación de los sistemas propuestos en esta competición que se está trabajando en un entorno donde

no hay acuerdo entre anotadores y que tanto los conjuntos de datos utilizados para crear los sistemas como las respuestas de estos deben contemplar este paradigma. Es por ello que la evaluación de dichos sistemas se debe realizar con herramientas que permitan comparar la anotación con desacuerdo del conjunto de datos de prueba (*ground truth*) con los resultados obtenidos (*system output*).

- **Verdad fundamental (*ground truth*)**: El desacuerdo entre anotadores se puede contemplar desde dos escenarios, uno estricto (*hard*) donde el valor de las diferentes anotaciones se debe simplificar a una única etiqueta (*gold*), por ejemplo, mediante el uso de voto mayoritario; o desde un escenario flexible (*soft*) donde es importante conservar el desacuerdo entre anotadores, por lo que la métrica de evaluación debe tener en cuenta la proporción de anotadores que han elegido cada etiqueta.

- **Salida del sistema (*system output*)**: Asimismo, en la salida de los sistemas se debe contemplar también un escenario estricto, donde los sistemas deben predecir una o varias categorías para cada instancia, y un escenario flexible, donde los sistemas deben predecir la probabilidad de cada categoría dentro de las posibles.

La combinación de los escenarios existentes, tanto desde la anotación como desde los resultados de los sistemas, ofrece tres tipos de evaluación:

- **Hard-hard**: en el que la etiqueta del *ground truth* y la de la salida del sistema deben coincidir.

- **Hard-soft**: en el que la etiqueta predicha por el sistema es comparada con la distribución de probabilidades de todas las clases en la verdad fundamental.

- **Soft-soft**: en el que la probabilidad de cada etiqueta definida por el sistema es comparada con la probabilidad de dicha etiqueta según la verdad fundamental.

En la sección 3.4 de este trabajo se detallan estos tres tipos de evaluación propuestos para la competición, mientras que en la sección 5.1.1 se describe la métrica utilizada para medir cada una de las evaluaciones mencionadas.

Finalmente, en la sección 5.5 se muestran los resultados obtenidos tanto en la fase de experimentación, evaluando el conjunto de datos de desarrollo y los resultados finales dentro de la competición, comparando con otros resultados obtenidos por otros participantes.

5.1.1. Métrica de evaluación - ICM

La selección de métricas idóneas para una tarea de clasificación dada, es un problema destacado en el ámbito de Procesamiento del Lenguaje Natural. Métricas como Accuracy, Precision, Recall o F-Measure, son comunes en este tipo de tareas, pero pueden diferir sustancialmente afectando seriamente el proceso de optimización de un sistema. Por ejemplo, si se asigna a todos los elementos, la clase mayoritaria sería muy efectivo según Accuracy y al mismo tiempo obtener un mal resultado según MAAC (Amigo y Delgado, 2022).

Para la evaluación de los sistemas propuestos en la competición EXIST 2023, se propone utilizar como métrica base de evaluación ICM (Information Contrast Model), una métrica basada en la similitud que generaliza PMI[1] y puede ser utilizada para evaluar los resultados de sistemas en problemas de clasificación calculando la similitud entre el resultado del sistema y el *gold standard*. La definición general de esta métrica es:

$$ICM(A, B) = \alpha_1 IC(A) + \alpha_2 IC(B) - \beta IC(A \cup B)$$

Donde $IC(A)$ es el Contenido de Información de la instancia (-log(P(A)) representada por el conjunto características de A (Amigo y Delgado, 2022). En el contexto de la competición, las categorías tienen una estructura jerárquica y las instancias pueden pertenecer a más de una categoría por lo que la métrica anterior se define como:

$$ICM(s(d), g(d)) = 2IC(s(d)) + 2IC(g(d)) - 3IC(s(d) \cup g(d))$$

Donde IC es el Contenido de Información, $s(d)$ el conjunto de etiquetas asignadas al documento d por el sistema s y $g(d)$ las etiquetas anotadas para el documento d en el *gold standard*.

[1] PMI (Pointwise Mutual Information) es una medida estadística que evalúa la relación entre dos eventos, calculando la probabilidad conjunta de que ambos eventos ocurran dividida por el producto de sus probabilidades individuales.

Por otro lado, hasta donde se sabe, no existe una métrica actual que permita evaluar problemas de clasificación jerárquica con múltiples etiquetas en un escenario de aprendizaje con desacuerdo, es por ello que desde la organización de la competición EXIST 2023 se ha planteado una extensión de esta métrica ICM llamada ICM-soft. ICM-soft es una métrica diseñada para aceptar valores basados en probabilidades tanto en la salida del sistema como en las anotaciones en el *gold standard*. Para hallar esta métrica, primero se define el Contenido de Información de una asignación única de una categoría c con un acuerdo v para una instancia en concreto.

$$I(\{\langle c, v\rangle\}) = -log_2(P(\{d \in D : g_c(d) \geq v\}))$$

Se puede observar que el contenido de información al asignar una categoría c con un acuerdo v disminuye a medida que la probabilidad de encontrar una instancia que reciba dicha categoría c con un acuerdo igual o mayor que v aumenta. Para abordar este aspecto, se realiza un cálculo utilizando la media y la desviación de los niveles de acuerdo para cada clase a través de las instancias, aplicando la probabilidad acumulativa sobre la distribución normal inferida. Además, en el caso de una varianza igual a cero (sin variabilidad en los niveles de acuerdo), se debe considerar que la probabilidad para valores iguales o inferiores a la media es 1 (contenido de información igual a cero), y la probabilidad para valores por encima de la media debe ajustarse (Plaza et al., 2023).

5.2. Experimentación

En esta sección se detallan los resultados obtenidos en la fase de desarrollo de los sistemas planteados para la competición EXIST 2023 para la detección de sexismo y la clasificación de la intención del autor al escribir el mensaje sexista en redes sociales. Estos resultados no se aportan directamente a la competición sino que permiten determinar qué subsistemas son más eficientes y cuáles no.

En el capítulo 4 de este trabajo, se desarrolla el sistema planteado para este trabajo, donde primero se realiza una evaluación de los grandes modelos del lenguaje existentes en el momento, frente a una tarea similar a la que abarca este trabajo, detección de sexismo con el conjunto de datos de EXIST 2021. Gracias a esta evaluación se seleccionan los modelos

PlanTL-GOB-ES/roberta-base-bne y **distilbert-base-uncased** para procesar los contenidos en español e inglés respectivamente. A continuación, se utilizan y se valoran diferentes estrategias de aumento de datos basadas en agregación de juicios y en repetición de etiquetado, tal como se puede observar en la tabla 4.4.

A partir de este conjunto de datos se han definido una serie de experimentos basados en tres estrategias de gestión del desacuerdo de las definidas en la publicación (Uma et al., 2021b):

- **Agregación de juicios *(judgement aggregation)***: en este planteamiento, descrito en 4.6.1, el desacuerdo entre anotadores se resuelve por voto mayoritario, donde en este caso al haber un número par de anotadores, el empate se resuelve definiendo la instancia en función de la clase mayoritaria para cada tarea.

- **Repetición de etiquetado *(repeated Labeling)***: este planteamiento descrito en la sección 4.6.2, se basa en la idea de que no tiene por qué haber una única etiqueta para cada caso, por lo que se debe intentar capturar el desacuerdo de los anotadores a partir de la distribución de etiquetas seleccionadas.

- **Aprendiendo de las masas *(learning from crowds)***: este planteamiento, descrito en la sección 4.6.3, se basa en las propias cohortes (tipos de etiquetadores por edad y género), de tal manera que se desarrolla un modelo para cada una y, a partir de los resultados se plantean diferentes opciones de agregación.

Finalmente, a partir de los conjuntos de datos desarrollados por aumento de datos con la base del conjunto EXIST 2023 aportado por la competición, de los dos modelos preentrenados seleccionados, uno para cada idioma, y de estas tres estratégicas, se han realizado más de 40 experimentos para la primera tarea y cerca de 30 para la segunda. Esta diferencia viene dada por un lado, porque algunas de las premisas válidas en la primera tarea por ser una tarea de clasificación binaria mono-etiqueta no se pueden dar en la segunda. Y por otro lado, porque algunas conclusiones extraídas se han aplicado en el desarrollo del sistema en la segunda, como por ejemplo el uso de algunos conjuntos de datos aumentados.

5.3. Resultados para la tarea 1 - Identificación de sexismo

En una evaluación previa a la competición, se han utilizado los conjunto de datos de entrenamiento y desarrollo como conjuntos de trabajo para entrenamiento y evaluación de la primera tarea. Se han evaluado más de 40 experimentos basados en las tres estrategias definidas, los modelos preentrenados seleccionados previamente y con configuraciones en las que se varía el conjunto de datos de entrenamiento y métodos de agregación, donde la respuesta inicial era múltiple. La tabla 5.1 muestra los mejores resultados obtenidos para cada una de las aproximaciones planteadas. A partir de estos resultados, se pueden extraer diferentes conclusiones. De un modo general y en una comparativa de las tres estrategias principales, se observa que la estrategia de obviar el desacuerdo entre anotadores (ag-3 & ag-7) es la que peores resultados ha dado de todo el espectro. Por otro lado, los diferentes sistemas basados en **aprendiendo de las masas** (*sys-lfc-**), donde se ha generado un modelo para cada cohorte, han generado mejores resultados en general, excepto en la evaluación en ICM-soft que el mejor resultado se obtiene con el sistema basado en **repetición de etiquetado** (rl-2 & rl-4).

En un análisis de cada uno de los resultados que se muestran en la tabla 5.1, se puede observar que:

ag-3 & ag-7: este resultado se obtiene con la combinación de dos modelos creados por mediante **agregación de juicios**, donde tal como se explica en el punto 4.6.1 de este trabajo, el desacuerdo entre anotadores se resuelve por voto mayoritario. Tal como se puede observar en la tabla 4.5, estos dos modelos están basados en un aumento de datos únicamente por traducción de los tuits en inglés a español para el ag-3, y los de español a inglés para el ag-7.

- El hecho de que estos dos modelos hayan dado mejores resultados que aquellos en los que se ha incluido el conjunto de datos de EXIST-2021, puede ser debido a que la variación en los anotadores perjudica al resultado final.

- Que se obtenga un mejor resultado en esta configuración de un modelo para cada idioma que en cualquiera de las otras en las que se combinan los idiomas (los modelos ag-9 a ag-16 de la tabla 4.5) es indicativo de que las expresiones de sexismo tienen sus peculiaridades

ID	Método	ICM hard-hard	F-measure hard-hard (YES)	F-Measure hard-hard (NO)	F-measure hard-hard (macro-F)	ICM hard-soft	ICM soft-soft
gold standard	gold standard hard	0.9995	1.0	1.0	1.0	1.2749	2.7257
ag-3 & ag-7	Agregación de datos (4.6.1)	0.5799	**0.8182**	0.8176	0.8152	0.3999	**0.8584**
rl-2 & rl-4	Repetición de etiquetado (4.6.2)	0.5873	0.8068	0.8272	0.817	0.4393	0.7447
sys-lfc-9	Aprendiendo de las masas: Voto mayoritario (4.6.3)	0.5841	0.8059	0.826	0.816	0.4563	-
sys-lfc-10	Aprendiendo de las masas: Voto basado en género; preferencia v. masculino (4.6.3)	0.5883	0.8091	0.8255	0.8173	0.4458	0.3913
sys-lfc-11	Aprendiendo de las masas: Voto basado en género; preferencia v. femenino (4.6.3)	**0.5942**	0.8105	0.8279	**0.8192**	**0.4758**	0.4125
sys-lfc-12	Aprendiendo de las masas: Voto basado en rangos de edad (4.6.3)	0.573	0.8	0.8246	0.8123	0.4214	-
sys-lfc-13	Aprendiendo de las masas: Distribución de probabilidad (media) (4.6.3)	0.575	0.8029	0.8233	0.8131	0.4306	0.7363
sys-lfc-14	Aprendiendo de las masas: Distribución de probabilidad (género) (4.6.3)	0.575	0.8029	0.8233	0.8131	0.4306	0.785
sys-lfc-15	Aprendiendo de las masas: Distribución de probabilidad (edad) (4.6.3)	0.5785	0.8037	0.8247	0.8142	0.4274	-1.5735
sys-lfc-16	Aprendiendo de las masas: Distribución de probabilidad (cohortes) (4.6.3)	0.5841	0.8038	**0.828**	0.8159	0.4485	0.8499

Tabla 5.1: Tarea 1: Esta tabla muestra los mejores resultados obtenidos en la evaluación de los sistemas propuestos utilizando el conjunto de datos de desarrollo como conjunto de evaluación. La primera columna muestra el *id* del sistema o modelo(s) utilizado. La columna método indica el método o estrategia que se ha seguido para obtener estos resultados, entre paréntesis se referencia al punto de este trabajo donde se describe el sistema. El resto de columnas contienen las métricas ICM con las tres configuraciones definidas anteriormente (hard-hard, hard-soft y soft-soft) así como las Métricas F-Measure de cada etiqueta y Macro-F de las dos etiquetas posibles.

según el idioma en que se exprese, modelos específicos de cada idioma tienden a capturar mejor estas expresiones y, por tanto, generar mejores resultados.

- Este sistema también se destaca en la métrica ICM soft-soft, la cual compara la distribución de probabilidad de las etiquetas con la verdad fundamental. La razón detrás de este destacado rendimiento podría ser que este sistema no procesa de ninguna manera el resultado obtenido por el modelo. En cambio, al trabajar directamente con la distribución de probabilidades, el modelo ha podido capturar de manera más efectiva la información sobre el desacuerdo entre anotadores y aplicarla en la predicción de nuevos tuits.

rl-2 & rl-4: el resultado obtenido en general por el sistema basado en **repetición de etiquetado** no es destacable para ninguna de las evalua-

ciones. Aun así, los resultados obtenidos en las evaluaciones *hard-hard* son indicativos de que puede ser un buen sistema a tener en cuenta en tareas de clasificación binaria.

sys-lfc-9: en la configuración de este sistema, el resultado se determina por voto mayoritario a partir de los resultados obtenidos por seis modelos que representan a las seis cohortes definidas para la competición, tal como se explica en el punto 4.6.3 de este trabajo. Si bien, igual que el sistema anterior, este no obtiene unos resultados destacables, dentro de los sistemas basados en **aprendizaje a partir de las masas** es uno de los sistemas con buenos resultados en la mayoría de las métricas. Es decir que dentro de estos sistemas, obtiene la 3 mejor posición en todas las evaluaciones *hard-hard* y la segunda posición para *hard-soft* tanto dentro de este conjunto como en general. Por tanto, es una configuración a tener en cuenta. **sys-lfc-10** y **sys-lfc-11**: la configuración de estos dos sistemas, basa su voto en el género dando prioridad a los modelos basados en cohortes masculinas el **sys-lfc-10** y cohortes femeninas el **sys-lfc-11**.

- Dado que estos dos sistemas en conjunto son los que han obtenido mejor resultado en una evaluación *hard-hard* es indicativo de que en una configuración de modelos basados en cohortes donde el voto se decide por género es acertada para tareas de clasificación binaria con desacuerdo entre anotadores.

- La diferencia entre estos dos modelos, constata que las anotaciones realizadas por mujeres tienen mayor sensibilidad en la detección de sexismo en general.

- Por otro lado, estos dos sistemas son los que obtienen peores resultados en una evaluación ICM *soft-soft*, esto puede deberse a que al basar el resultado únicamente por las etiquetas escogidas por cada modelo, se pierde información importante respecto a la probabilidad de cada etiqueta. Esto lo demuestra el hecho de que el sistema *sys-lfc-14* que es similar, pero basado en probabilidades, obtiene el tercer mejor resultado en la misma evaluación.

sys-lfc-12: la configuración de este sistema es similar a la de los dos anteriores (*sys-lfc-10* y *sys-lfc-11*) pero en este caso el voto se realiza en función de los tres grupos de edad, tal como se explica en el punto 4.6.3. Este sistema obtiene los peores resultados en casi todas las evaluaciones, lo cual es

indicativo de que el rango de edad no afecta en cuanto a la identificación de sexismo. **sys-lfc-13**, **sys-lfc-14** y **sys-lfc-15**: la configuración de estos sistemas se basa en la agregación de la probabilidad de cada etiqueta obtenida por los modelos basados en cohortes. El primero haciendo la media de cada etiqueta, el segundo comparando los valores obtenidos según el género y el tercero, comparando los valores agrupando por la edad de las cohortes en la que se basan los modelos.

- Estos 3 sistemas obtienen algunos de los peores resultados de esta lista para las evaluaciones *hard-hard* y el sistema *sys-lfc-15* obtiene el peor resultado en ICM *soft-soft* con amplia diferencia, lo cual es indicativo de que esta aproximación no es idónea para tareas clasificación en las que haya desacuerdo.
- Por otro lado, el sistema *sys-lfc-14* obtiene el 3 mejor resultado en la evaluación *ICM soft-soft*, lo cual reafirma la importancia que tienen los modelos basados en el género para detectar sexismo.

sys-lfc-16: este sistema retorna la etiqueta con mayor probabilidad de los seis modelos basados en cohortes.

- Es el sistema con mejor resultado en la evaluación de la clase negativa (*F-Measure hard-hard(NO)*). Este resultado, es además, muy similar al obtenido por *sys-lfc-11*, lo cual se podria explicar si en muchos casos, la etiqueta más probable es una inferida por un modelo basado en una cohorte femenina.
- Por otro lado, también obtiene el segundo mejor resultado en la evaluación ICM *soft-soft*, y el tercelo en la evaluación ICM *hard-soft* lo cual indica que es un sistema a tener en cuenta en tareas de detección de sexismo con desacuerdo.

5.4. Resultados en tarea 2 - Identificación de la intención del autor

Al igual que para la primera tarea, para la segunda, también se ha realizado una experimentación previa utilizando el conjunto de datos de desarrollo como conjunto de evaluación. En esta tarea, es una tarea de clasificación jerárquica multi-clase, donde la jerarquía viene dada por aquellas instancias

seleccionadas como sexistas en las que se debe determinar la intención de la fuente. Para resolver esta jerarquía, se ha utilizado un sistema en cascada donde la decisión de sí un tuit es sexista o no la toma el sistema sys-lfc-16 y las instancias sexistas se evalúan con el resto de propuestas aquí planteadas. Siendo el resultado de la evaluación de la etiqueta NO **0.828** en todos los casos.

ID modelo	**Método**	**F-measure hard-hard (JUDGEMENTAL)**	**F-Measure hard-hard (REPORTED)**	**F-Measure hard-hard (DIRECT)**	**F-Measure hard-hard (macro-F)**
gold standard	gold standard hard	1.0	1.0	1.0	1.0
sys-lfc-9	Aprendiendo de las masas: Voto mayoritario(4.6.3)	**0.3178**	0.3394	0.5945	0.5199
sys-lfc-10	Aprendiendo de las masas: Voto basado en género(4.6.3)	0.2812	0.3293	0.5951	0.5084
sys-lfc-12	Aprendiendo de las masas: Voto basado en rangos de edad(4.6.3)	0.3015	0.3509	**0.6083**	0.5222
sys-lfc-13	Aprendiendo de las masas: Distribución de probabilidad (media)(4.6.3)	0.2974	**0.3659**	0.5988	**0.5225**
sys-lfc-14	Aprendiendo de las masas: Distribución de probabilidad (género)(4.6.3)	0.3099	0.358	0.5937	0.5224
sys-lfc-15	Aprendiendo de las masas: Distribución de probabilidad (edad)(4.6.3)	0.2588	0.3294	0.5922	0.5021

Tabla 5.2: Tarea 2: Esta tabla muestra los mejores resultados obtenidos en la evaluación de los sistemas propuestos utilizando el conjunto de datos de desarrollo como conjunto de evaluación. La primera columna muestra el *id* del sistema o modelo(s) utilizado. La columna método indica el método o estrategia que se ha seguido para obtener estos resultados, entre paréntesis se referencia al punto de este trabajo donde se describe el sistema. El resto de columnas contienen las Métricas F-Measure de cada etiqueta y Macro-F de las dos etiquetas posibles.

Las tablas 5.2 y 5.3 muestran los mejores resultados de los experimentos realizados para esta tarea. En un análisis de estos resultados obtenidos para la segunda tarea usando el conjunto de datos de desarrollo como conjunto de datos de evaluación, se puede observar que:

sys-lfc-10: la configuración de este sistema, se basa en el género de los modelos implementados a partir de la anotación de las cohortes. Esta configuración para la tarea 2, no obtiene tan buenos resultados como para la tarea 1, lo cual puede indicar que en este caso el género no influye tanto en el momento de determinar la intención del autor de un mensaje sexista.

sys-lfc-12: la configuración de este sistema, se basa en la edad de los modelos implementados a partir de la anotación de las cohortes. Pese a que no obtiene tampoco los mejores resultados, si que son significativamente mejores a los del sistema anterior, lo cual indica que la edad es más determinante que el género de una persona para indicar la intención del autor.

sys-lfc-13: este sistema se basa en la agregación de la probabilidad de cada etiqueta obtenida por los modelos basados en cohortes, haciendo media

ID modelo	Method	**ICM hard-hard**	**ICM hard-soft**	**ICM soft-soft**
gold standard	gold standard hard	1.5989	-3.1749	4.6802
sys-lfc-9	Aprendiendo de las masas: Voto mayoritario(4.6.3)	0.2936	-6.8743	-
sys-lfc-10	Aprendiendo de las masas: Voto basado en género(4.6.3)	0.2811	-6.6572	-
sys-lfc-12	Aprendiendo de las masas: Voto basado en rangos de edad(4.6.3)	0.3106	-6.7823	-
sys-lfc-13	Aprendiendo de las masas: Distribución de probabilidad (media)(4.6.3)	**0.3142**	-6.5681	-1.2093
sys-lfc-14	Aprendiendo de las masas: Distribución de probabilidad (género)(4.6.3)	0.3021	-6.8162	**-1.1792**
sys-lfc-15	Aprendiendo de las masas: Distribución de probabilidad (edad)(4.6.3)	0.279	**-6.4868**	-2.2366

Tabla 5.3: Tarea 2: Esta tabla muestra los mejores resultados obtenidos en la evaluación de los sistemas propuestos utilizando el conjunto de datos de desarrollo como conjunto de evaluación. La primera columna muestra el *id* del sistema o modelo(s) utilizado. La columna método indica el método o estrategia que se ha seguido para obtener estos resultados, entre paréntesis se referencia al punto de este trabajo donde se describe el sistema. El resto de columnas contienen las métricas ICM con las tres configuraciones definidas anteriormente (hard-hard, hard-soft y soft-soft).

por cada etiqueta para obtener su resultado, tal como se explica en el punto 4.6.3.

- En esta segunda tarea obtiene el mejor resultado en la evaluación ICM *hard-hard*, lo cual indica que es un buen sistema para tareas de clasificación multi-clase.

- También obtiene su mejor resultado en la métrica *F-Measure hard-hard (macro-F)*, lo cual indica que es un sistema que en general ha obtenido buenos resultados en la clasificación de todas las etiquetas.

sys-lfc14: este sistema se basa en la agregación de la probabilidad de cada etiqueta obtenida por los modelos basados en cohortes y comparando los resultados por género. Obtiene el mejor resultado en la evaluación ICM *soft-soft*, lo cual indica en este caso que el género de los anotadores sí que influye en la detección de la intención del autor de un mensaje sexista.

5.5. Resultados finales

En este apartado se presentan las estrategias utilizadas finalmente para cada una de las ejecuciones presentadas en las tareas 1 y 2 de la competición.

Para la primera tarea se han seleccionado un sistema basado en **repetición de etiquetado** (*rl-2 & rl-4*) dado que en los resultados de la experimentación que se muestra en el apartado 5.3 se observa que su evaluación de las métricas basadas en ICM obtiene un mejor rendimiento en general a pesar de que el sistema basado en agregación de juicios tenga mejor resultado en ICM soft-soft su rendimiento en las otras métricas es peor. Como segundo sistema se propone el basado en **aprendiendo de las masas** *(sys-lfc-11)* con agregación por género con preferencia para el voto de modelos basados en cohorte femenina, por ser el que más métricas ha obtenido con mejores resultados. Finalmente, como tercer sistema para la primera tarea se propone también uno basado en **aprendiendo de las masas**, pero en este caso en agregación por distribución de probabilidad por cohorte *(sys-lfc-16)* al ser el que mejor resultado obtiene en la métrica ICM soft-soft después del de agregación.

Para la segunda tarea se han seleccionado todos los sistemas basados en **aprendiendo de las masas** dado que son los con los que más se ha experimentado y que mejores resultados se han obtenido en la experimentación. El primero, el primero basado en voto mayoritario *(sys-lfc-9)* al tener en general buenos resultados en las métricas basadas en *F-Measure*. El segundo basado en media de distribución de probabilidad *(sys-lfc-13)* al ser el que en experimentación ha obtenido mejores resultados en la métrica ICM hard-hard. Para la tercera ejecución se propone el sistema basado en distribución de probabilidad por género, dado que es el que genera mejor resultado en la métrica ICM soft-soft.

La tabla 5.4 proporciona un resumen de las estrategias empleadas en cada una de las ejecuciones que han sido finalmente presentadas para las tareas 1 y 2 de la competición EXIST 2023.

Los resultados de las ejecuciones para las tareas 1 y 2 se presentan en las tablas 5.5, 5.6, 5.7, 5.8, 5.9 y 5.10. Cada tabla proporciona detalles sobre varias métricas de evaluación para español, inglés, y la combinación de ambos idiomas. La primera columna dentro de cada grupo de columnas indica la posición de clasificación de cada ejecución. Las dos primeras filas de valores en cada tabla representan los resultados de referencia (0) basado en el *gold* y

ID	Tarea	Ejecución	Método
rl-2 & rl-4	Tarea 1	1	Repetición de etiquetado
sys-lfc-11	Tarea 1	2	Voto basado en género; preferencia v. femenino
sys-lfc-16	Tarea 1	3	Aprendiendo de las masas: Distribución de probabilidad (cohortes)
sys-lfc-9	Tarea 2	1	Aprendiendo de las masas: voto mayoritario
sys-lfc-13	Tarea 2	2	Aprendiendo de las masas: Distribución de probabilidad (media)
sys-lfc-14	Tarea 2	3	Aprendiendo de las masas: Distribución de probabilidad (género)

Tabla 5.4: Ejecuciones enviadas para las tareas 1 y 2 de EXIST 2023: Esta tabla muestra el listado de ejecuciones enviadas para la tarea 1 y tarea 2 de EXIST 2023. En la primera columna se muestra el identificador de sistema seleccionado, en la segunda la tarea para la que se ha aplicado, en la tercera el número de ejecución asignado y en la última el método en el que se basa el sistema.

los mejores resultados obtenidos (1 o 2 si el resultado no ha superado alguna línea base) para la tarea respectiva en cada evaluación.

5.5.1. Tarea 1 - detección de sexismo

En la tarea 1 de la competición EXIST 2023, participaron 24 equipos con aproximadamente 70 envíos para la primera tarea, 40 en la segunda y 38 en la tercera. En las tablas 5.5, 5.6 y 5.7 se muestran los resultados obtenidos para las evaluaciones *hard-hard*, *hard-soft* y *soft-soft* de la tarea 1 respectivamente. Las tres columnas principales muestran los resultados agrupados por idioma, siendo la primera resultados en una evaluación bilingüe (español-inglés), la segunda únicamente los resultados obtenidos de evaluar únicamente las instancias en español, y finalmente, la tercera columna para los resultados obtenidos de evaluar únicamente las instancias en inglés. En los resultados de cada una de las evaluaciones, se aporta una línea de *gold standard* con el mejor resultado posible, es decir, aquel en el que la salida del sistema es igual al *gold standard* como medida de sistema imbatible.

Ejecución	Bilingüe				ES				EN			
	Posición	ICM-Hard	ICM-Hard Norm	F1	Posición	ICM-Hard	ICM-Hard Norm	F1	Posición	ICM-Hard	ICM-Hard Norm	F1
Gold	0	0.9948	1	1		0.9999	1	1		0.9798	1	1
Mejor resultado	1	0.6575	0.785	0.8109	1	0.6995	0.8011	0.8384	1	0.6004	0.7693	0.776
JPM_UNED_1	28	0.5057	0.6883	0.756	20	0.514	0.6783	0.7748	30	0.4819	0.6972	0.7308
JPM_UNED_2	33	0.4863	0.6759	0.7533	24	0.5016	0.6701	0.7784	39	0.4556	0.6812	0.7204
JPM_UNED_3	19	**0.5223**	**0.6989**	**0.7623**	14	**0.545**	**0.6988**	**0.7885**	29	**0.4844**	**0.6987**	**0.7284**

Tabla 5.5: Resultados para la evaluación hard-hard de la Tarea 1: Esta tabla presenta los resultados obtenidos en la evaluación hard-hard. Las tres columnas principales representan los resultados para diferentes idiomas: la primera columna muestra los resultados para todas las instancias (bilingüe), la segunda columna se centra en instancias en español (es), y la tercera columna se centra en instancias en inglés (en). Cada columna incluye las siguientes métricas para cada ejecución: ICM, ICM normalizado y Medida F1 de la clase positiva (YES). La primera fila indica el resultado del estándar de referencia, mientras que la segunda fila representa el mejor resultado. Las filas restantes muestran los resultados obtenidos por nuestras tres ejecuciones de evaluación.

Tal como se puede observar, el mejor resultado obtenido teniendo en cuenta todas las instancias, ha sido la posición diecinueve con un resultado de **0.6989** en la métrica ICM-Hard Norm obtenida con la tercera ejecución de las enviadas, la basada en **prendiendo de las masas: Distribución de probabilidad (género)** con ID **sys-lfc-14**. El mejor resultado en esta evaluación para esta tarea, ha sido el obtenido por el equipo **Mario**(Tian, Huang, y Zhang, 2023) con su tercera ejecución, obteniendo un resultado de **0.785**. Esta misma ejecución, evaluando únicamente las instancias en español, se ha obtenido un resultado en la métrica ICM-hard Norm de **0.6988**, quedando en la posición catorce. En esta evaluación el mejor resultado ha sido también para el equipo **Mario** (Tian, Huang, y Zhang, 2023) en este caso con su primera ejecución. Finalmente, para la evaluación de las instancias en inglés, el mejor resultado obtenido en la métrica ICM-hard Norm ha sido **0.6987** también con la tercera ejecución. El mejor resultado en esta evaluación ha sido **0.6004** obtenida por el equipo **SINAI** con su primera ejecución(Vallecillo-Rodríguez et al., 2023).

El sistema presentado por el equipo **Mario** se basa en el uso del modelo "GPT-NeoX" de código abierto para inglés y el modelo "BERTIN-GPT-J-6B" para español, entrenando algunos modelos en el dominio de detección de sexismo y otros se ajustaron utilizando conjuntos de entrenamiento basados en discurso de odio. A partir de estos resultados se deduce que el modelo preentrenado "BERTIN-GPT-J-6B" es un buen candidato a tener en cuenta para tareas similares en español con una evaluación hard-hard. Por otro lado, el sistema presentado por **SINAI** se basa en el modelo **mDeBERTa**, en técnicas de aumento de datos por repetición de etiquetado y en incorporación de información categórica de los anotadores (género y edad) en el proceso de entrenamiento. El hecho de que este sistema destaque en la evaluación *hard-hard* teniendo en cuenta únicamente las instancias en inglés, es demostrativo de que más que el sistema, el modelo utilizado tiene un buen rendimiento para este caso.

En la tabla 5.6, se muestran los resultados obtenidos para la primera tarea con una evaluación *hard-soft* donde las tres columnas principales referencian las evaluaciones: la primera contiene todas las instancias (bilingüe), la segunda únicamente con las instancias en español y la tercera con únicamente las instancias en inglés. Tal como se puede observar, de las tres ejecuciones enviadas, la tercera es la que tiene mejor posición en las tres

Ejecución	Bilingüe			ES			EN		
	Posición	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm
Gold	0	3.1182	1	0	3.1177	1	0	3.1141	1
Mejor resultado	2	0.4719	0.5725	2	0.6998	0.5752	2	0.1402	0.5709
JPM_UNED_1	19	0.1685	0.5235	17	0.3485	0.5135	31	-0.1243	0.5327
JPM_UNED_2	32	0.1075	0.5136	23	0.2927	0.5037	38	-0.1879	0.5235
JPM_UNED_3	17	**0.2041**	**0.5292**	13	**0.3927**	**0.5212**	28	**-0.0924**	**0.5373**

Tabla 5.6: Resultados para la evaluación hard-soft de la Tarea 1: Esta tabla presenta los resultados obtenidos en la evaluación hard-soft. Las tres columnas principales representan los resultados para diferentes idiomas: la primera columna muestra los resultados para todas las instancias, la segunda columna se centra en instancias en español y la tercera columna se centra en instancias en inglés. Cada columna incluye las siguientes métricas para cada ejecución: ICM-Soft e ICM-Soft normalizado. La primera fila indica el resultado del estándar de referencia, mientras que la segunda fila representa el mejor resultado. Las filas restantes muestran los resultados obtenidos por nuestras tres ejecuciones de evaluación.

evaluaciones de esta tabla. Con la posición diecisiete en la evaluación bilingüe con un resultado de **0.5292** en la métrica ICM-Soft Norm y el mejor resultado en esta evaluación es del equipo **Mario** con su tercera ejecución que obtiene un resultado de 0.6987 en la métrica ICM-Soft Norm. En la evaluación en español, se obtiene la treceava posición con un resultado de **0.5212** en la métrica ICM-Soft Norm y la mejor posición es del equipo **Mario** con un resultado de 0.5852 para la misma métrica. Finalmente, en la evaluación de las instancias en inglés, se obtiene la posición 28 con un resultado de 0.5373 en la métrica ICM-Soft Norm y la mejor posición es para el equipo **SINAI** con un resultado de 0.5709.

Se observa en esta tabla que los mejores resultados son los obtenidos por la línea base **EXIST2023_oracle_most_voted** que considera la etiqueta más votada por los anotadores y alcanza una puntuación de 0.6897 en este contexto. Esto seguramente se deba a que para hacer la evaluación *hard-soft* se convierte la distribución de probabilidad de la salida del sistema a etiquetas, perdiendo gran parte de la información del desacuerdo, lo cual habrá afectado a la evaluación de los sistemas.

Los resultados de la evaluación *soft-soft* de la tarea 1, se muestran en la tabla 5.7. En esta tabla, se puede observar que de las tres ejecuciones enviadas, la primera es la que tiene un mejor rendimiento en el contexto bilingüe, clasificándose en la doceava posición con un resultado de **0.6058** en la métrica ICM-Soft Norm, y también en el contexto de evaluación de instancias en inglés obteniendo un resultado de **0.6463** en la métrica ICM-Soft Norm. En cambio, en el contexto de instancias en español, la ejecución de las enviadas con mejor rendimiento es la segunda, que obtiene un resultado de **0.5689** en

RUN	Bilingüe				ES				EN			
	Ejecución	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Gold	0	3.1182	1	0.5472	0	3.1177	1	0.5208	0	3.1141	1	0.577
Mejor resultado	1	0.903	0.6421	0.796	1	0.9527	0.6196	0.7672	1	0.8157	0.6683	1.0198
JPM_UNED_1	12	**0.6779**	**0.6058**	**0.8023**	18	0.6536	0.5671	**0.7588**	11	**0.6632**	**0.6463**	0.8512
JPM_UNED_2	18	0.5972	0.5927	**0.8852**	16	**0.6641**	**0.5689**	0.8116	15	0.4853	0.6207	0.9677
JPM_UNED_3	29	0.2467	0.5361	2.2342	31	0.1576	0.4799	2.1581	24	0.3506	0.6012	2.3196

Tabla 5.7: Resultados para la evaluación soft-soft de la Tarea 1: Esta tabla presenta los resultados obtenidos en la evaluación soft-soft. Las tres columnas principales representan los resultados para diferentes idiomas: la primera columna muestra los resultados para todas las instancias, la segunda columna se centra en instancias en español y la tercera columna se centra en instancias en inglés. Cada columna incluye las siguientes métricas para cada ejecución: ICM-Soft, ICM-Soft normalizado y Entropía Cruzada. La primera fila indica el resultado del estándar de referencia, mientras que la segunda fila representa el mejor resultado. Las filas restantes muestran los resultados obtenidos por nuestras tres ejecuciones de evaluación.

la métrica ICM-Soft Norm. Por otro lado, la ejecución con mejor resultado, tanto en el contexto bilingüe como en el contexto de instancias en español, es la tercera de las enviadas por el equipo **SINAI** (Vallecillo-Rodríguez et al., 2023), que obtiene un resultado de 0.6421 en la métrica ICM-Soft Norm en el contexto bilingüe y un resultado de 0.6196 en el contexto en español. En el contexto en inglés; sin embargo, es el equipo **CLassifiers** el que obtiene la primera posición con un resultado de 0.6683 en la métrica ICM Soft Norm. El planteamiento de este equipo para esta tarea, se basa en el modelo CardiffNLP Twitter XLM-RoBERTa con aumento de datos con el conjunto de datos de EXIST 2021(Ersoy, Radler, y Carpentieri, 2023).

5.5.2. Tarea 2 - Identificación de intención del emisor

Los resultados obtenidos para la tarea 2 (categorizar los tuits según la intención de la fuente) se muestran en las Tablas 5.8, 5.9 y 5.10. En una evaluación general de esta tarea, el sistema que destaca es el de la segunda ejecución, basado en **Aprendiendo de las masas: Distribución de probabilidad (media)** con ID **sys-lfc-13**, que obtiene la segunda y tercera posiciones en la evaluación *soft-soft* para las instancias en español y bilingüe respectivamente.

Los resultados obtenidos en la tarea dos en una evaluación *hard-hard*, se muestran en la tabla, 5.8, donde se puede observar que el sistema basado en **Aprendiendo de las masas: Distribución de probabilidad (media)** obtiene las posiciones ocho y nueve en las evaluaciones de contexto español y bilingüe respectivamente con un resultado de **0.6992** en la primera y de **0.712** en la segunda, ambos resultados con la métrica de ICM-hard Norm. En el contexto de instancias en inglés, se obtiene la posición once con un

Ejecución	Bilingüe				ES				EN			
	Posición	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1
Gold	0	1.5378	1	1	0	1.6007	1	1	0	1.4449	1	1
Mejor resultado	1	0.4887	0.7764	0.5715	1	0.5711	0.7732	0.6059	1	0.3677	0.781	0.5224
JPM_UNED_1	11	0.1673	0.7079	0.5032	10	0.1986	0.6911	0.5281	12	0.1024	0.727	0.4661
JPM_UNED_2	9	**0.1862**	**0.712**	0.5054	8	**0.2351**	**0.6992**	0.5341	13	0.0995	0.7264	0.4649
JPM_UNED_3	10	0.1806	0.7108	**0.5092**	9	0.2231	0.6965	**0.5383**	11	**0.1034**	**0.7272**	**0.4673**

Tabla 5.8: Resultados para la evaluación hard-hard de la Tarea 2: Esta tabla presenta los resultados obtenidos en la evaluación hard-hard. Las tres columnas principales representan los resultados para diferentes idiomas: la primera columna muestra los resultados para todas las instancias, la segunda columna se centra en instancias en español y la tercera columna se centra en instancias en inglés. Cada columna incluye las siguientes métricas para cada ejecución: ICM, ICM normalizado y Medida F. La primera fila indica el resultado del estándar de referencia, mientras que la segunda fila representa el mejor resultado. Las filas restantes muestran los resultados obtenidos por nuestras tres ejecuciones de evaluación.

resultado 0.7272 en la métrica ICM-hard Norm. En esta evaluación, los mejores resultados de los 3 contextos son para el equipo **Mario** que obtiene un resultado de 0.7764 en la evaluación bilingüe, de 0.7732 en la evaluación de contexto en español y 0.781 en el contexto inglés, en todos los casos con la métrica ICM-hard Norm.

Los resultados obtenidos en esta evaluación, junto a los obtenidos por el equipo de Mario, indican que la aproximación basada en **aprendiendo de las masas** puede ser una buena aproximación para tareas con desacuerdo, basando esta estrategia en los modelos "GPT-NeoX" para inglés y "BERTIN-GPT-J-6B" para español.

Ejecución	Bilingüe			ES			EN		
	Posición	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm
Gold	0	6.2057	1	0	6.2431	1	0	6.1178	1
EXIST2023_oracle_most_voted	1	-2.3974	0.7803	1	-1.8502	0.7684	1	-3.265	0.7943
Mejor resultado	2	-5.12	0.7108	2	-4.693	0.6871	3	-5.8008	0.7387
JPM_UNED_1	23	-7.5078	0.6498	21	-6.8073	0.6266	25	-8.9034	0.6707
JPM_UNED_2	22	**-7.3346**	**0.6542**	19	**-6.5622**	**0.6336**	23	**-8.8583**	**0.6717**
JPM_UNED_3	24	-7.5205	0.6495	22	-6.8533	0.6253	24	-8.8978	0.6708

Tabla 5.9: Resultados para la evaluación hard-soft de la Tarea 2: Esta tabla presenta los resultados obtenidos en la evaluación hard-soft. Las tres columnas principales representan los resultados para diferentes idiomas: la primera columna muestra los resultados para todas las instancias, la segunda columna se centra en instancias en español y la tercera columna se centra en instancias en inglés. Cada columna incluye las siguientes métricas para cada ejecución: ICM-Soft e ICM-Soft normalizado. La primera fila indica el resultado del estándar de referencia, mientras que la segunda fila representa el mejor resultado. Las filas restantes muestran los resultados obtenidos por nuestras tres ejecuciones de evaluación.

En la tabla 5.9 se puede observar como el sistema basado en **Aprendiendo de las masas: Distribución de probabilidad (media)** obtiene los mejores resultados de entre los enviados a la competición, pero los otros dos sistemas propuestos están en posiciones consecutivas a este en casi todos los contextos. En esta evaluación se observa que ningún sistema ha podido

superar la línea base **EXIST2023_oracle_most_voted**, basada en escoger directamente la etiqueta más votada. Esto puede deberse a la perdida de información en la evaluación *hard-soft* que afecta en general a todos los sistemas. En esta evaluación, las ejecuciones que han obtenido mejores resultados son: i)el equipo **UMUTeam** (García-Díaz, Pan, y Valencia-García, 2023) que obtiene la posición dos con un resultado de 0.7108 en el contexto bilingüe, su sistema de este equipo se basa en el uso de una combinación de diferentes LLMs con Funciones de Etiqueta (LFs) y entrenarlos con datos multilingües; ii) el equipo **SMS** que obtuvo la segunda posición en el contexto en español con un resultado de 0.6871 en la métrica ICM-Soft Norm, su sistema se basa principalmente en el modelo GloVe (Global Vectors for Word Representation) con una dimensionalidad de 200 para los textos en inglés y incrustaciones de palabras FastText de 300 dimensiones para textos en español, junto con una capa LSTM; iii) el equipo **UMUTeam** obtuvo la tercera posición en el contexto en inglés dado que en este contexto la segunda posición la ocupa el baseline **EXIST2023_test_majority_class** sistema basado en que todas las instancias tengan la clase mayoritaria.

Ejecución	Bilingüe				ES				EN			
	Posición	ICM-Soft	ICM-Hard Soft	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Gold	0	6.2057	1	0.9128	0	6.2431	1	0.8926	0	6.1178	1	0.9354
Mejor resultado	1	-1.3443	0.8072	1.7833	1	-1.2317	0.7861	1.6415	1	-1.1471	0.8407	1.8001
JPM_UNED_2	3	**-1.675**	**0.7988**	**2.5549**	2	**-1.4414**	**0.7801**	**2.4472**	6	-2.1062	0.8197	2.6757
JPM_UNED_3	5	-1.6888	0.7984	2.5561	3	-1.5006	0.7785	2.4511	4	**-2.0436**	**0.8211**	**2.674**

Tabla 5.10: Resultados para la evaluación soft-soft de la Tarea 2: Esta tabla presenta los resultados obtenidos en la evaluación soft-soft. Las tres columnas principales representan los resultados para diferentes idiomas: la primera columna muestra los resultados para todas las instancias, la segunda columna se centra en instancias en español y la tercera columna se centra en instancias en inglés. Cada columna incluye las siguientes métricas para cada ejecución: ICM-Soft, ICM-Soft normalizado y Entropía Cruzada. La primera fila indica el resultado del estándar de referencia, mientras que la segunda fila representa el mejor resultado. Las filas restantes muestran los resultados obtenidos por nuestras tres ejecuciones de evaluación.

Los resultados de la evaluación *soft-soft* de la tarea 2, se muestran en la tabla 5.10. En esta tabla, se puede observar que de las dos ejecuciones enviadas, la basada en **Aprendiendo de las masas: Distribución de probabilidad (media)** es la que obteiene mejores resultados, alcanzando la segunda y tercera posición en los contextos en español con un resultado **0.7988** y bilingüe con un resultado **0.7801** respectivamente. En cambio, para el contexto en inglés es el sistema basado en **Aprendiendo de las masas: Distribución de probabilidad (género)** que obtiene la cuarta posición. En esta evaluación obtiene el mejor resultado el equipo **DRIM**(Erbani et

al., 2023) en el contexto bilingüe con un resultado de 0.8072 en la métrica ICM-Hard Soft y en el contexto en inglés con un resultado de 0.8407 en la métrica ICM-Soft Norm. Este equipo basa su estrategia en la creación de tres modelos a partir del LLMs BERT, uno para cada tarea, concatenados a los que se les añade información como la longitud del tuit y el número de hastags. En el contexto en español, el equipo que obtiene mejor posición es **AIT_FHSTP**(Böck et al., 2023), que obtiene un resultado de 0.7861 en la métrica ICM-Soft Norm; su propuesta se centra en el uso de *embeddings* derivados de un modelo de análisis de sentimientos y un modelo de análisis de toxicidad, los cuales a su vez se basan en XLM_RoBERTa.

Capítulo 6

Conclusiones y trabajo futuro

Este trabajo se centra en la identificación de sexismo en redes sociales en un contexto donde el conjunto de datos conserva el desacuerdo entre anotadores. Para ello se ha participado en la edición de 2023 de la competición EXIST en el marco del CLEF 2023[1].

Este capítulo recopila las diferentes conclusiones extraídas del trabajo realizado, y propone algunas líneas de trabajo futuro con el fin de mejorar los resultados obtenidos y alcanzar una mayor comprensión del efecto que tiene el desacuerdo entre anotadores tanto en la detección de sexismo como de cualquier otro contenido tóxico.

6.1. Conclusiones

El crecimiento exponencial de la web y las redes sociales en los últimos años ha favorecido la comunicación entre usuarios a nivel global, permitiendo a gente de todo el plantea estar en contacto entre sí, compartir intereses y mantener relaciones de forma independiente a la distancia, lo que ha contribuido significativamente a la globalización de la sociedad y la cultura. Por otro lado, ya sea por la oportunidad que ofrece el anonimato o por la impersonalidad de estas plataformas o por otra causa, en estos entornos ha proliferado el uso del lenguaje tóxico de odio tanto de manera general como hacia personas y/o grupos concretos. Estos comportamientos abusivos

[1] https://clef2023.clef-initiative.eu/

afecta no solo a los usuarios que las sufren, sino también a las propias redes sociales, que ven como la proliferación de estos comportamientos producen un abandono masivo por parte de los usuarios. Es por ello que tanto agentes públicos como privados e investigadores abogan cada año por iniciativas para el control de lenguaje de odio en redes sociales. Entre estos, destaca el colectivo femenino, que cada vez es más objeto de acoso y discursos de odio en plataformas como las redes sociales. Este fenómeno, conocido como sexismo, se caracteriza por manifestaciones lingüísticas de odio o prejuicio hacia las mujeres, que incluyen formas de exclusión social, discriminación, hostilidad, amenazas violentas y objetivación sexual.

Con el análisis del estado del arte en este campo se han repasado diferentes iniciativas en torno a la detección de lenguaje tóxico en general incluyendo las ediciones de 2021 y 2022 de la competición EXIST de las que se han analizado los diferentes enfoques que han realizado los participantes a estas. En general, se ha estudiado una evolución de las estrategias utilizadas durante los últimos años, se ha comprobado que antes de la aparición de la tecnología transformer, los enfoques se centraban en el uso de técnicas tradicionales de machine learning tales como SVM o RF, y la remarcable diferencia en los resultados que obtienen los sistemas que se basan en una u otra.

En la edición de EXIST 2023 en la que se basa este trabajo, las tareas se basan en el paradigma de *Learning With Disagreements (LeWiDi)* o aprendizaje con desacuerdo. Este se basa en la idea de que en la forma tradicional de anotar un conjunto de datos (realizar una anotación por diferentes anotadores expertos en la materia y en los casos de desacuerdo entre estos consensuar una etiqueta común ya sea por voto mayoritario o por otro sistema) se pierde información, por lo que es necesario trabajar con conjuntos de datos que conserven el desacuerdo entre anotadores y definir estrategias para abordarlo. Para ello se crea un conjunto de datos anotado por diferentes anotadores de 6 cohortes definidas por su género y su edad.

Estudiando la bibliografía en torno al paradigma de *LeWiDi*, se han observado diferentes estrategias con las que abordar este problema, para el desarrollo de este trabajo se han hecho tres planteamientos: el primero basado en la premisa de que el desacuerdo no existe y agrupar las diferentes anotaciones por voto mayoritario; el segundo basado en intentar capturar el desacuerdo de los anotadores dentro del propio modelo del lenguaje y,

el último entrenando un modelo para cada cohorte de las utilizadas en la anotación. Destaca el hecho que los sistemas en los que se ha tenido en cuenta el género de los anotadores para definir la etiqueta han obtenido en general buenos resultados, sobre todo cuando, en caso de empate, se ha resuelto a partir de los resultados generados por los modelos basados en cohorte femenina.

De los tres planteamientos realizados para la competición se extrae que:

En el primer sistema, basado en agregación de juicios, el sistema realizado depende directamente de la capacidad de cómputo, en sistemas observados de otros participantes, se observa que utilizando modelos preentrenados más grandes y con aumento de datos, se obtienen finalmente, mejores en una evaluación *hard*.

En el segundo sistema, basado en repetición de etiquetado, la premisa es intentar introducir el desacuerdo de los usuarios en el propio modelo. Se observan resultados no destacables que quizás convendría volver a plantear con variaciones en el aumento de datos o con diferentes modelos preentrenados.

En el tercer sistema, basado en aprendiendo de las masas, ha sido el único planteamiento presentado a la competición basado en información socio-demográfica de los anotadores. Las variantes de este planteamiento han generado diferentes resultados que confirman la importancia que tiene el sesgo de cada anotador.

Por otro lado, analizando los resultados de la competición se observa que diferentes equipos participantes han incluido en su sistema un aumento de datos basado en la edición de 2021 de EXIST obteniendo mejores resultados que los propios. Esto indica que puede que se haya cometido un error en la experimentación en este planteamiento.

Finalmente, de los resultados obtenidos en la competición y en comparación a los sistemas que mejores resultados han obtenido en cada evaluación, se ha observado la importancia del uso de modelos bilingües frente al uso de modelos monolingües. También se observan buenos resultados en el uso de la información socio-demográfica al implementar seis modelos basados en las seis cohortes para calcular la distribución final de probabilidades.

6.2. Trabajo futuro

En vista de los resultados obtenidos en la competición EXIST, se puede suponer que el planteamiento utilizado para abordar el desacuerdo entre anotadores es correcto, pero por otro lado, también se observa que varios de los sistemas que han obtenido mejores resultados han utilizado el conjunto de datos de EXIST 2021 como aumento de datos, si bien es cierto que estos sistemas no abordan directamente el desacuerdo entre anotadores, sería conveniente volver a tener en cuenta esta estrategia de aumento de datos.

También se ha observado el uso de otros modelos preentrenados que no se han tenido en cuenta en la selección inicial utilizada para el desarrollo de este trabajo, por lo que sería interesante abordar por ejemplo los modelos "BERTIN-GPT-J-6B" y "GPT-NEOX" utilizados por el grupo Mario y observar su rendimiento frente a los utilizados.

Finalmente, de las estrategias existentes para abordar el paradigma de desacuerdo entre anotadores, sería interesante profundizar en el sistema utilizado en este trabajo basado en la estrategia de "aprendizaje de las masas", dado que ha dado un buen resultado y permite, además, tener en cuenta el punto de vista de usuarios concretos en función de su perfil socio-demográfico en el momento de valorar si un contenido es o no ofensivo.

Bibliografía

Bibliografía

[Ahuir2022] Ahuir, Vicent. 2022. Enhancing Sexism Identification and Categorization in Low-data Situations.

[Akiwowo et al.2020] Akiwowo, Seyi, Bertie Vidgen, Vinodkumar Prabhakaran, y Zeerak Waseem, editores. 2020. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Online, Noviembre. Association for Computational Linguistics.

[Albarqouni et al.2016] Albarqouni, Shadi, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, y Nassir Navab. 2016. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321, May.

[Amigo y Delgado2022] Amigo, Enrique y Agustín Delgado. 2022. Evaluating extreme hierarchical multi-label classification. En Smaranda Muresan Preslav Nakov, y Aline Villavicencio, editores, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 5809–5819, Dublin, Ireland, Mayo. Association for Computational Linguistics.

[Anzovino, Fersini, y Rosso2018] Anzovino, Maria, Elisabetta Fersini, y Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. *Lecture Notes in Computer Science*, página 57–64, Jan.

[Aroyo et al.2019] Aroyo, Lora, Lucas Dixon, Nithum Thain, Olivia Redfield, y Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study

of understanding toxicity in online discussions. *Companion Proceedings of The 2019 World Wide Web Conference*, May.

[Aroyo y Welty2015] Aroyo, Lora y Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *Ai Magazine*, 36(1):15–24, Mar.

[Bahdanau, Cho, y Bengio2014] Bahdanau, Dzmitry, Kyunghyun Cho, y Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

[Basile et al.2019a] Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, y Manuela Sanguinetti. 2019a. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. En *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 54–63, Minneapolis, Minnesota, USA, Junio. Association for Computational Linguistics.

[Basile et al.2019b] Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, y Manuela Sanguinetti. 2019b. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. En Jonathan May Ekaterina Shutova Aurelie Herbelot Xiaodan Zhu Marianna Apidianaki, y Saif M. Mohammad, editores, *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 54–63, Minneapolis, Minnesota, USA, Junio. Association for Computational Linguistics.

[Basile et al.2021] Basile, Valerio, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, y others. 2021. We need to consider disagreement in evaluation. En *Proceedings of the 1st workshop on benchmarking: past, present and future*, páginas 15–21. Association for Computational Linguistics.

[Beigman y Beigman Klebanov2009] Beigman, Eyal y Beata Beigman Klebanov. 2009. Learning with annotation noise. En Keh-Yih Su Jian Su Janyce Wiebe, y Haizhou Li, editores, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International*

Joint Conference on Natural Language Processing of the AFNLP, páginas 280–287, Suntec, Singapore, Agosto. Association for Computational Linguistics.

[Beigman Klebanov, Beigman, y Diermeier2008] Beigman Klebanov, Beata, Eyal Beigman, y Daniel Diermeier. 2008. Analyzing disagreements. En Ron Artstein Gemma Boleda Frank Keller, y Sabine Schulte im Walde, editores, *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, páginas 2–7, Manchester, UK, Agosto. Coling 2008 Organizing Committee.

[Bengoetxea2022] Bengoetxea, Kepa. 2022. Multiaztertest@Exist-Iberlef2022: Sexism Identification in Social Networks.

[Biber1988] Biber, Douglas. 1988. *Variation across Speech and Writing.* Cambridge University Press.

[Biber1989] Biber, Douglas. 1989. A typology of english texts.

[Böck et al.2023] Böck, Jaqueline, Mina Schütz, Daria Liakhovets, Nathanya Queby Satriani, Andreas Babic, Djordje Slijepčević, Matthias Zeppelzauer, y Alexander Schindler. 2023. Ait_fhstp at exist 2023 benchmark: sexism detection by transfer learning, sentiment and toxicity embeddings and hand-crafted features. *Working Notes of CLEF.*

[Caselli et al.2020] Caselli, Tommaso, Valerio Basile, Jelena Mitrović, y Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english.

[Cañete2019] Cañete, José. 2019. Compilation of large spanish unannotated corpora. May.

[Cañete et al.2020a] Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, y Jorge Pérez. 2020a. Spanish pre-trained bert model and evaluation data. En *PML4DC at ICLR 2020.*

[Cañete et al.2020b] Cañete, José, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, y Jorge Pérez. 2020b. Spanish pre-trained bert model and evaluation data. En *PML4DC at ICLR 2020.*

[Clarke y Grieve2017] Clarke, Isobelle y Jack Grieve. 2017. Dimensions of abusive language on Twitter. En Zeerak Waseem Wendy Hui Kyong Chung Dirk Hovy, y Joel Tetreault, editores, *Proceedings of the First Workshop on Abusive Language Online*, páginas 1–10, Vancouver, BC, Canada, Agosto. Association for Computational Linguistics.

[Conneau et al.2019] Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, y Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

[Daniel et al.2018] Daniel, Florian, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, y Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing. *ACM Computing Surveys*, 51(1):1–40, Jan.

[Davani et al.2023] Davani, Aida Mostafazadeh, Mohammad Atari, Brendan Kennedy, y Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319, 03.

[Dawid y Skene1979] Dawid, Alexander Philip y Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

[de Paula, da Silva, y Schlicht2021] de Paula, Angel Felipe Magnossão, Roberto Fray da Silva, y Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. *arXiv preprint arXiv:2111.04551*.

[Devlin et al.2018a] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, y Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding.

[Devlin et al.2018b] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, y Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding.

[Devlin et al.2018c] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, y Kristina Toutanova. 2018c. Bert: Pre-training of deep bidirectional transformers for language understanding.

[Duggan2017] Duggan, Maeve. 2017. Online Harassment 2017, Julio.

[Dumitrache, Aroyo, y Welty2019] Dumitrache, Anca, Lora Aroyo, y Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. En Jill Burstein Christy Doran, y Thamar Solorio, editores, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 2164–2170, Minneapolis, Minnesota, Junio. Association for Computational Linguistics.

[Dumitrache et al.2018] Dumitrache, Anca, Oana Inel, Lora Aroyo, Benjamin Timmermans, y Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement.

[Erbani et al.2023] Erbani, Johan, Előd Egyed-Zsigmond, Diana Nurbakova, y Pierre-Edouard Portier. 2023. When multiple perspectives and an optimization process lead to better performance, an automatic sexism identification on social media with pretrained transformers in a soft label context. *Working Notes of CLEF*.

[Ersoy, Radler, y Carpentieri2023] Ersoy, Berna Ilke, Gian Radler, y Sofia Carpentieri. 2023. Classifiers at exist 2023: Detecting sexism in spanish and english tweets with xlm-t.

[Fandiño et al.2022a] Fandiño, Asier Gutiérrez, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, y Marta Villegas. 2022a. Maria: Spanish language models.

[Fandiño et al.2022b] Fandiño, Asier Gutiérrez, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, y Marta Villegas. 2022b. Maria: Spanish language models.

[Fersini et al.2022] Fersini, Elisabetta, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, y Jeffrey Sorensen. 2022. *SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification.*

[Fersini, Nozza, y Rosso2020] Fersini, Elisabetta, Debora Nozza, y Paolo Rosso. 2020. AMI @ EVALITA2020: Automatic Misogyny Identification. En Valerio Basile Danilo Croce Maria Maro, y Lucia C. Passaro, editores, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*. Accademia University Press, páginas 21–28.

[Fišer et al.2018] Fišer, Darja, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, y Jacqueline Wernimont, editores. 2018. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, Octubre. Association for Computational Linguistics.

[García-Díaz, Pan, y Valencia-García2023] García-Díaz, José Antonio, Ronghao Pan, y Rafael Valencia-García. 2023. Umuteam at exist 2023: Sexism identification and categorisation fine-tuning multilingual large language models.

[Gimpel et al.2011] Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, y Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. En Dekang Lin Yuji Matsumoto, y Rada Mihalcea, editores, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, páginas 42–47, Portland, Oregon, USA, Junio. Association for Computational Linguistics.

[Guan et al.2017] Guan, Melody Y, Varun Gulshan, Andrew M Dai, y Geoffrey E Hinton. 2017. Who said what: Modeling individual labelers improves classification.

[Gutiérrez-Fandiño et al.] Gutiérrez-Fandiño, Asier, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, y Marta Villegas. *MarIA: Spanish Language Models MarIA: Modelos del Lenguaje en Español.*

[Hartmann et al.2023] Hartmann, Jochen, Mark Heitmann, Christian Siebert, y Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, Mar.

[He, Gao, y Chen2021] He, Pengcheng, Jianfeng Gao, y Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

[Hosseini et al.2017] Hosseini, Hossein, Sreeram Kannan, Baosen Zhang, y Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments, Febrero. arXiv:1702.08138 [cs].

[Hung et al.2013] Hung, Tam Nguyen, Lam Ngoc Tran, y Karl Aberer. 2013. An evaluation of aggregation techniques in crowdsourcing, Oct.

[Inel et al.2014] Inel, Oana, Khalid Khamkham, T Cristea, Anca Dumitrache, Arne Rutjes, Jelle Ploeg, Łukasz Romaszko, Lora Aroyo, y Robert-Jan Sips. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. *Springer eBooks*, página 486–504, Jan.

[Jamison y Gurevych2015] Jamison, Emily y Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. En Lluís Màrquez Chris Callison-Burch, y Jian Su, editores, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, páginas 291–297, Lisbon, Portugal, Septiembre. Association for Computational Linguistics.

[Javier et al.2022] Javier, Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo Gonzalez, Manu Romero, y Marıa Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling.

[Klebanov y Beigman2009] Klebanov, Beata Beigman y Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503, Dec.

[Krizhevsky, Hinton, y others2009] Krizhevsky, Alex, Geoffrey Hinton, y others. 2009. Learning multiple layers of features from tiny images.

[Kumar et al.2020] Kumar, Ritesh, Atul Kr. Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock, y Daniel Kadar, editores. 2020. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France, Mayo. European Language Resources Association (ELRA).

[Kumar et al.2018a] Kumar, Ritesh, Atul Kr. Ojha, Shervin Malmasi, y Marcos Zampieri. 2018a. Benchmarking Aggression Identification in Social Media. En Ritesh Kumar Atul Kr. Ojha Marcos Zampieri, y Shervin Malmasi, editores, *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, páginas 1–11, Santa Fe, New Mexico, USA, Agosto. Association for Computational Linguistics.

[Kumar et al.2018b] Kumar, Ritesh, Atul Kr. Ojha, Marcos Zampieri, y Shervin Malmasi, editores. 2018b. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, Santa Fe, New Mexico, USA, Agosto. Association for Computational Linguistics.

[Kumar et al.2022] Kumar, Ritesh, Atul Kr. Ojha, Marcos Zampieri, Shervin Malmasi, y Daniel Kadar, editores. 2022. *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, Gyeongju, Republic of Korea, Octubre. Association for Computational Linguistics.

[la Rosa et al.2022] la Rosa, Javier De, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, y María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling.

[Lalor, Wu, y Yu2019] Lalor, John P., Hao Wu, y Hong Yu. 2019. Soft label memorization-generalization for natural language inference.

[Lan et al.2019] Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, y Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

[Liu et al.2019a] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, y Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach.

[Liu et al.2019b] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, y Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

[Nguyen, Vu, y Nguyen2020] Nguyen, Dat Quoc, Thanh Vu, y Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. En *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, páginas 9–14.

[Nobata et al.2016] Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, y Yi Chang. 2016. Abusive Language Detection in Online User Content. En *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, páginas 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Paun et al.2018] Paun, Silviu, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, y Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

[Pérez et al.2022] Pérez, Juan Manuel, Damián Ariel Furman, Laura Alonso Alemany, y Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. En Nicoletta Calzolari Frédéric Béchet Philippe Blache Khalid Choukri Christopher Cieri Thierry Declerck Sara Goggi Hitoshi Isahara Bente Maegaard Joseph Mariani Hélène Mazo Jan Odijk, y Stelios Piperidis, editores, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, páginas 7235–7243, Marseille, France, Junio. European Language Resources Association.

[Peterson et al.2019a] Peterson, Joshua, Ruairidh Battleday, Thomas Griffiths, y Olga Russakovsky. 2019a. Human uncertainty makes classification more robust. En *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, páginas 9616–9625.

[Peterson et al.2019b] Peterson, Joshua C, Ruairidh M Battleday, Thomas L Griffiths, y Olga Russakovsky. 2019b. Human uncertainty makes classification more robust.

[Plank, Hovy, y Søgaard2014] Plank, Barbara, Dirk Hovy, y Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. En Shuly Wintner Sharon Goldwater, y Stefan Riezler, editores,

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, páginas 742–751, Gothenburg, Sweden, Abril. Association for Computational Linguistics.

[Plaza et al.2023] Plaza, Laura, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, y Paolo Rosso. 2023. Overview of exist 2023–learning with disagreement for sexism identification and characterization. En *International Conference of the Cross-Language Evaluation Forum for European Languages*, páginas 316–342. Springer.

[Plaza-del Arco et al.2021] Plaza-del Arco, Flor Miriam, M Dolores Molina-González, LAU López, y MT Martín-Valdivia. 2021. Sexism identification in social networks using a multi-task learning system. En *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain*, volumen 2943, páginas 491–499.

[Poesio y Artstein2005] Poesio, Massimo y Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. En Adam Meyers, editor, *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, páginas 76–83, Ann Arbor, Michigan, Junio. Association for Computational Linguistics.

[Poesio et al.2019] Poesio, Massimo, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, y Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. En Jill Burstein Christy Doran, y Thamar Solorio, editores, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 1778–1789, Minneapolis, Minnesota, Junio. Association for Computational Linguistics.

[Rangel et al.2021] Rangel, Francisco, Gretel Liz de la Peña-Sarracén, María Alberta Chulvi-Ferriols, Elisabetta Fersini, y Paolo Rosso. 2021. Profiling Hate Speech Spreaders on Twitter Task at PAN 2021. páginas 1772–1789.

[Raykar et al.2010] Raykar, Vikas C., Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, y Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322.

[Reidsma y Carletta2008] Reidsma, Dennis y Jean Carletta. 2008. Squibs: Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

[Reidsma y op den Akker2008] Reidsma, Dennis y Rieks op den Akker. 2008. Exploiting 'subjective' annotations. En Ron Artstein Gemma Boleda Frank Keller, y Sabine Schulte im Walde, editores, *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, páginas 8–16, Manchester, UK, Agosto. Coling 2008 Organizing Committee.

[Risch y Krestel2020] Risch, Julian y Ralf Krestel. 2020. Toxic comment detection in online discussions. *Algorithms for intelligent systems*, página 85–109, Jan.

[Roberts et al.2019] Roberts, Sarah T., Joel Tetreault, Vinodkumar Prabhakaran, y Zeerak Waseem, editores. 2019. *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, Agosto. Association for Computational Linguistics.

[Rodrigues y Pereira2018] Rodrigues, Filipe y Francisco C Pereira. 2018. Deep learning from crowds. *Proceedings of the ... AAAI Conference on Artificial Intelligence*, 32(1), Apr.

[Rodríguez-Sánchez, Carrillo-de Albornoz, y Plaza2020] Rodríguez-Sánchez, Francisco, Jorge Carrillo-de Albornoz, y Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

[Rodríguez-Sánchez et al.2021] Rodríguez-Sánchez, Francisco, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, y Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0):195–207.

[Rodríguez-Sánchez et al.2022] Rodríguez-Sánchez, Francisco, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, y Paolo Rosso. 2022. Overview of exist 2022: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 69(0):229–240.

[Russell et al.2007] Russell, Bryan, Antonio Torralba, Kevin Murphy, y William T Freeman. 2007. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, Oct.

[Schütz et al.] Schütz, M, J Boeck, D Liakhovets, D Slijepcevic, A Kirchknopf, M Hecht, J Bogensperger, S Schlarb, A Schindler, y M Zeppelzauer. Automatic sexism detection with multilingual transformer models, corr abs/2106.04908 (2021). *URL: https://arxiv. org/abs/2106.04908*.

[Shao et al.2022] Shao, Zhou, Ruoyan Zhao, Sha Yuan, Ming Ding, y Yongli Wang. 2022. Tracing the evolution of ai in the past decade and forecasting the emerging trends. *Expert Systems with Applications*, 209:118221–118221, Dec.

[Sharmanska et al.2016] Sharmanska, Viktoriia, Daniel Hernandez-Lobato, Jose Miguel Hernandez-Lobato, y Novi Quadrianto. 2016. Ambiguity helps: Classification with disagreements in crowdsourced annotations. *Thecvf.com*, página 2194–2202.

[Sheng, Provost, y Ipeirotis2008] Sheng, Victor, Foster J Provost, y Panos Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers, Aug.

[Sheshadri y Lease2013] Sheshadri, Aashish y Matthew Lease. 2013. Square: A benchmark for research on computing crowd consensus. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 1:156–164, Nov.

[Sheth, Shalin, y Kursuncu2022] Sheth, Amit, Valerie L. Shalin, y Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318, Junio.

[Simpson et al.2019] Simpson, Edwin, Erik-Lân Do Dinh, Tristan Miller, y Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. En Anna Korhonen David Traum, y Lluís Màrquez, editores, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 5716–5728, Florence, Italy, Julio. Association for Computational Linguistics.

[Sutskever, Vinyals, y Le2014] Sutskever, Ilya, Oriol Vinyals, y Quoc V Le. 2014. Sequence to sequence learning with neural networks.

[Swim et al.2001] Swim, Janet K, Lauri L Hyers, Laurie Cohen, y Melissa J Ferguson. 2001. Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social Issues*, 57(1):31–53, Jan.

[Tamayo y Bueno] Tamayo, Roberto Labadie y Reynier Ortega Bueno. Are Examples Worth More Than Language?

[Taulé Delor et al.2021] Taulé Delor, Mariona, Alejandro Ariza, Montserrat Nofre, Enrique Amigó, y Paolo Rosso. 2021. Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish. *Procesamiento del lenguaje natural, 2021, num. 67, p. 209-221.*

[Tian, Huang, y Zhang2023] Tian, Lin, Nannan Huang, y Xiuzhen Zhang. 2023. Efficient multilingual sexism detection via large language models cascades. *Working Notes of CLEF.*

[Uma et al.2021a] Uma, Alexandra, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, y Massimo Poesio. 2021a. SemEval-2021 Task 12: Learning with Disagreements. En *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, páginas 338–347, Online, Agosto. Association for Computational Linguistics.

[Uma et al.2021b] Uma, Alexandra, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, y Massimo Poesio. 2021b. Learning from Disagreement: A Survey | Journal of Artificial Intelligence Research.

[Vaca-Serrano2022] Vaca-Serrano, Alejandro. 2022. Detecting and classifying sexism by ensembling transformers models. *language*, 2:1.

[Vallecillo-Rodríguez et al.2023] Vallecillo-Rodríguez, María Estrella, FMP del Arco, Luis Alfonso Ureña-López, María Teresa Martín-Valdivia, y Arturo Montejo-Ráez. 2023. Integrating annotator information in transformer fine-tuning for sexism detection. *Working Notes of CLEF*.

[Vaswani et al.2017] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, y Illia Polosukhin. 2017. Attention is All you Need. En I. Guyon U. Von Luxburg S. Bengio H. Wallach R. Fergus S. Vishwanathan, y R. Garnett, editores, *Advances in Neural Information Processing Systems*, volumen 30. Curran Associates, Inc.

[Victor Sanh2023a] Victor Sanh, Lysandre Debut, Julien Chaumond Thomas Wolf (Hugging Face). 2023a. Jun.

[Victor Sanh2023b] Victor Sanh, Lysandre Debut, Julien Chaumond Thomas Wolf (Hugging Face). 2023b. Jun.

[Vogels2021] Vogels, Emily A. 2021. Feb.

[Vogels2022] Vogels, Emily A. 2022. Teens and cyberbullying 2022, Dec.

[Waseem et al.2017a] Waseem, Zeerak, Wendy Hui Kyong Chung, Dirk Hovy, y Joel Tetreault. 2017a. proceedings of the first workshop on abusive language online (alw) - acl anthology$_2$017.

[Waseem et al.2017b] Waseem, Zeerak, Thomas Davidson, Dana Warmsley, y Ingmar Weber. 2017b. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

[Wiegand, Siegel, y Ruppenhofer2018] Wiegand, Michael, Melanie Siegel, y Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. *Oeaw.ac.at*, página 1–10.

[Zampieri et al.2019] Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, y Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). En *Proceedings of the 13th International Workshop on Semantic Evaluation*, páginas 75–86.

[Zampieri et al.2020] Zampieri, Marcos, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, y Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). En *Proceedings of SemEval*.

[Álvarez et al.2022] Álvarez, Victoria Pachón, Jacinto Mata Vázquez, Wissam Chibane, y Juan Luis Domínguez. 2022. Automatic Sexism Identification Using an Ensemble of Pretrained Transformers.

Apéndice A

Combining Transformer Based Language Models with Socio-demographic Information for Improving Sexism Detection in Social Media

Combining Transformer Based Language Models with Socio-demographic Information for Improving Sexism Detection in Social Media

Notebook for the EXIST Lab at CLEF 2023

Jacobo Pedrosa-Marín[1,*], Jorge Carrillo-de-Albornoz[1,2] and Laura Plaza[1,2]

[1]NLP & IR Group, Universidad Nacional de Educación a Distancia (UNED), 28040, Spain
[2]RMIT University, 3000, Australia

Abstract

Detecting and addressing sexism in social networks is crucial for fostering inclusive and respectful digital spaces. The third edition of the EXIST competition emphasizes the importance of incorporating annotator disagreement into the classification process, recognizing the inherent challenges and diversity of perspectives in identifying sexist content. In this paper, we present our participation in the EXIST 2023 campaign where we propose systems for Task 1 (Sexism Identification) and task 2 (Source Intention Identification), both for hard and soft evaluation contexts. We adopted a primary strategy that involved data augmentation to enhance the training dataset. By leveraging techniques such as translation and the use of transformers, we aimed to expand the available data and capture a broader range of linguistic patterns and expressions related to sexism. Additionally, the EXIST 2023 dataset allows to identify and exploit annotators characteristics such as gender and age. We have used this socio-demographic information to train different models that capture each age-gender cohort singularities, and used different strategies to combine them in a final decision, in the hard approaches, and a probability representation, in the soft approaches. The results achieved suggest that having different models for the different cohorts improves the efficiency of the classification.

Keywords

Sexism Detection, Sexism Identification Learning with disagreement, Transformer Models, Natural Language Processing

1. Introduction

In recent years, the rise of social media platforms such as Twitter and Facebook has brought about a significant transformation in communication and society. These platforms have provided users with new means to express their ideas, thoughts, and knowledge. These platforms hold tremendous potential for information dissemination, and researchers have extensively examined their impact in various fields, including politics and medicine. However, the proliferation of hate speech on these platforms has emerged as a growing concern. The rise of hate speech

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
jacobo.pedrosa@lsi.uned.es (J. Pedrosa-Marín); jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz); lplaza@lsi.uned.es (L. Plaza)
0009-0004-1224-3760 (J. Pedrosa-Marín)

poses a significant challenge, demanding careful attention and effective solutions to ensure a safe and inclusive online environment [1].

Detecting and preventing hate speech in social medial can be challenging, especially considering the overwhelming volume of data generated on these platforms every second, which makes it necessary to employ automated methods and advanced technologies to process and classify the content efficiently. Over the years, numerous studies and competitions have emerged, focusing on the analysis and automation of online community management. These initiatives aim to address various challenges associated with content detection and moderation, including Anomaly Detection [2], Phishing Detection [3], Toxicity Detection [4] and Sexism Detection [5, 6].

In this paper, we focus on a particular form of harmful content: sexist expressions. Sexism encompasses actions or attitudes that exhibit prejudice or discrimination towards individuals based on their gender. It is closely linked to societal beliefs and expectations regarding the roles that individuals should adhere to, with its repercussions primarily impacting women. Detecting sexism in online platforms is crucial for creating inclusive and respectful digital spaces. It enables the identification and moderation of harmful content, promotes gender equality, and helps to prevent the perpetuation of discriminatory behaviors.

The EXIST challenges serves as an avenue for researchers and practitioners to develop and present their approaches and models in tackling the complex task of sexism detection in social media [7]. The 2021 and 2022 editions of the competition were held in the IberLEF forum [1] and were the first shared tasks on sexism detection in social networks whose aim was to identify and classify various forms of sexism, ranging from explicit and hostile expressions to more nuanced or even benevolent behaviors involving implicit sexism. With participation from over 50 teams from research institutions and companies worldwide, the substantial interest shown by the research community underscores the significance of the problem at hand.

The third edition of the EXIST challenge at CLEF builds upon the tasks addressed in previous years, while facing a new challenge: the identification of the author's intention behind sexist messages. However, the main innovation is the adoption of the "learning with disagreements" paradigm [8] and the participation of annotators of different genders and ages. This approach aims to mitigate the potential "label bias" by incorporating diverse perspectives and sensitivities from different population groups, ensuring a more comprehensive reflection of viewpoints and recognizing that annotators' socio-demographic backgrounds can shape their labeling decisions [9, 10].

This paper presents the participation of the JPM-UNED team in EXIST 2023 at CLEF. Our approach includes the use of transformers along with different strategies (such as voting and aggregation) to leverage the collective knowledge and the disagreement among the annotators to derive the most reliable predictions. By considering the socio-demographic variances among annotators and employing tailored models and strategies, our objective is to enhance the accuracy and robustness of our classification approach. This approach contributes to a more deeper understanding of the complexities involved in sexism detection and enhances the overall effectiveness of our model.

This paper is organized as follows: Section 2 reviews the state of the art in sexism detection

[1] https://sites.google.com/view/iberlef2022/home

and learning with disagreements; Section 3 presents this edition of EXIST, including an overview of the three tasks proposed and the dataset provided; Section 4 presents the systems developed for participating in the competition; Section 5 discusses the results; and Section 6 summarizes the main conclusions and discusses potential improvements for future work.

2. Related Work

In this section, we first examine the importance of sexism detection in social networks, review previous research in this area, and provide an overview of previous editions of the EXIST competition. We then briefly discuss the state-of-the-art in Learning With Disagreements (**LeWiDi**) to consider different approaches for managing tasks of this nature.

The detection of sexism has traditionally been regarded as a distinct form of hate speech [11]. It can be approached through various methods, such as data-driven models that incorporate n-grams and additional features [12], classical machine learning models [13], and deep learning models that utilize LSTM and CNN architectures [14]. Additionally, certain studies have explored the utilization of offensive lexicons like Hurtlex [15].

However, it is important to recognize that sexism is not always expressed as hate speech. As highlighted in [7], sexism can take on a "friendly" or even "humorous" tone, such as in the case of benevolent sexism and sexist jokes [16]. Consequently, novel approaches are necessary to detect the various forms of sexism, ranging from hostile and explicit to subtle and seemingly benign expressions.

In the following section, we present the EXIST 2021 and 2022 editions, which introduced the challenging task of detecting sexism in all its nuanced manifestations.

2.1. The EXIST Challenges

In the previous two editions of the EXIST competition, two tasks were proposed. Task 1 focused on sexism identification, aiming to detect whether a given post contains sexist content or not. Task 2, on the other hand, focused on sexism categorization, aiming to classify the type of sexism present in a post into one of the following five classes: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence y (v) misogyny and non-sexual violence.

In the 2021 edition of EXIST, the majority of submissions for both tasks relied on transformer-based models for classification. Out of the 23 participating teams, 14 teams utilized BERT, a widely used transformer model, as the foundation of their solutions. Additionally, 10 teams employed BETO, a version of BERT that is trained on Spanish text. Furthermore, 5 teams leveraged XLM-R, a multilingual variant of RoBERTa that supports multiple languages, including Spanish [5]. In the 2022 edition of EXIST, all participating teams utilized transformer-based solutions. Among these solutions, 8 teams used BERT, 5 teams used BETO, and 4 teams used RoBERTa. The widespread adoption of transformer-based models in both editions of EXIST highlights their effectiveness in tackling the challenges associated with sexism detection in social networks [6].

2.2. Working with Disagreements

Usually, in the fields of Artificial Intelligence (AI) and Natural Language Processing, datasets are built with instances having a single class or interpretation referred to as the "gold standard". However, this approach fails to capture the nuances of human behavior, which often involves disagreement and varying perspectives. Tasks that involve subjectivity or ambiguity inherently introduce the possibility of biased annotations influenced by the perspectives of individual annotators. Moreover, socio-demographic factors, such as education, age or gender, have the potential to influence the annotation process and introduce biases.

An emerging solution that is gaining popularity is to engage multiple annotators, ideally representing diverse demographic strata and socioeconomic contexts, and retain the labels provided by each annotator instead of relying solely on a gold standard. This approach allows the systems to incorporate various perspectives for each instance, enabling them to learn from different points of view.

Working with a dataset that lacks a unanimous label for each instance offers valuable insights, but it also requires ways to manage the divergent opinions among annotators. The current state of the art in learning with disagreement can be categorized into four main categories:

- **Judgements aggregation**: Methods that operate under the assumption that only a single "truth" exists for each instance typically aggregate all crowd annotations into a single label, commonly referred to as the "silver" label. There are multiple approaches to tackle this challenge, with the most straightforward being the adoption of a "majority vote". However, one of the widely employed techniques is the utilization of **Probabilistic aggregation methods**, which leverage the probabilities assigned to each label by individual annotators [17, 18, 19].
- **Filtering hard items:** this approach utilizes the disagreement information to filter the dataset by removing instances with significant disagreement. Within this category, [20] proposed two approaches. The first approach involves directly discarding instances that exhibit disagreement among annotators. The second approach is to train separate models for each annotator and discard predictions that demonstrate substantial disagreement.
- **Learning directly from crowds:** This classification approach acknowledges the absence of a single truth or gold standard and instead focuses on training a classifier directly from the crowd, utilizing probabilistic distributions or soft labels. The objective is to capture the collective knowledge of the annotators and incorporate their diverse perspectives into a single model. There are various strategies to approach this. For instance, [21] propose a "repeated labeling method" where replicas of each instance are created for each label, enabling multiple annotations per instance. Another interesting method in this category is presented by [22], which involves adding a crowd layer after the output layer during training.
- **Using both hard labels and information about disagreements:** These methods utilize both gold labels and disagreement to train the models. One approach is to train with hard labels, while incorporating crowd information as part of the loss function during the training process. This allows the model to learn from both the ground truth labels and the disagreements among the crowd annotations, improving its performance and capturing the collective knowledge of the annotators [23].

As we can see, there are various approaches available to address the challenge of working with disagreements. For further information on this topic, please refer to the survey conducted by [23]. This survey provides more in-depth insights and details on different methods and techniques that can be employed to handle disagreement in various tasks or domains.

3. The EXIST 2023 Lab at CLEF 2023

3.1. EXIST 2023 Tasks

The EXIST 2023 edition proposes the following three tasks: (i) sexism detection, (ii) source intention classification, and (iii) sexism categorization (see [10] for a detailed description). For each task, participants may provide both hard (a single "gold" label) and soft (a probabilistic label) outputs.

- **Task 1 - Sexism Detection:** The first task is a binary classification task where systems must decide whether or not a given tweet is sexist.
- **Task 2 - Source Intention Classification:** This task aims to categorize the message according to the intention of the author, which provides insights in the role played by social networks in the emission and dissemination of sexist messages. In this task, we propose a ternary classification task: (i) direct sexist message, (ii) reported sexist message and (iii) judgmental message.
- **Task 3 - Sexism Categorization:** Each sexist tweet must be categorized in one or more of the following categories, that reflect the facets of a woman's life that are the focus of the sexist message: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence.

3.2. EXIST 2023 Dataset

The EXIST 2023 dataset comprises tweets in both English and Spanish. The training set consists of over 3,200 tweets per language, while the development set includes 500 tweets per language. Additionally, the test set contains 1,000 tweets per language. To ensure diverse perspectives and mitigate label bias, each tweet in the dataset has been annotated by six individuals recruited through the Prolific service [2]. The annotators' gender (male/female[3]) and age (18-22 years old, 23-45 years old, +46 years old) are taken into account during the labeling process. Consequently, each tweet is labeled by annotators from a different gender and a different age groups.

For a more comprehensive understanding of the EXIST 2023 dataset, please refer to [10].

4. System description

In this section, we present our systems developed for participating in Task 1 and Task 2 of the EXIST 2023 competition. Our approach builds upon the methods discussed in the previous

[2] https://app.prolific.co

[3] Only male and female genders were consider for availability reasons

section and utilizes Transformer pre-trained models. To establish a baseline, we first examine the models that have been employed in previous editions of the competition. We then train and test our models using the EXIST 2021 Dataset for each language, employing the same configuration. The results for Spanish tweets are presented in Table 1, while Table 2 showcases the results for English tweets.

Table 1
Models review Spanish - baselines

Model	F-Measure
xlm-roberta-base	0.703
xlm-roberta-large	0.447
bert-base-multilingual-cased	0.701
distilbert-base-multilingual-cased	0.706
PlanTL-GOB-ES/roberta-base-bne	**0.751**
PlanTL-GOB-ES/roberta-large-bne	0.741
bertin-project/bertin-roberta-base-spanish	0.747
dccuchile/bert-base-spanish-wwm-cased	0.709
CenIA/distillbert-base-spanish-uncased	0.713

Table 2
Models review English - baselines

Model	F-Measure
xlm-roberta-base	0.652
xlm-roberta-large	0.676
bert-base-multilingual-cased	0.721
distilbert-base-multilingual-cased	0.701
roberta-large	0.288
distilbert-base-uncased	**0.741**
bert-base-uncased	0.733

After analyzing the results of our initial experiments (see Tables 1 and 2, we have selected the model with the best performance in Spanish, which is **PlanTL-GOB-ES/roberta-base-bne**, and in English, which is **distilbert-base-uncased**. Our remaining experiments will be conducted using these two models as the foundation. Based on these findings, we have decided to utilize these two models as the base for our future experiments.

As a previous step, we have also created an augmented version of the EXIST 2023 dataset by translating tweets from English to Spanish and vice versa. With this expanded dataset in hand, we have two parallel approaches to enhance the training process and our analysis.

Our initial approach involved fine-tuning the pre-trained models we had selected using the "repeated labeling" technique. This technique involved converting each instance in the datasets into six duplicated instances, with each instance assigned a unique label representing a single annotator. We will refer to this approach as "Learning from raw disagreement".

In our second approach, we trained six different models for each language, corresponding

to **each cohort based on age-gender combinations**. We utilized the individual votes of the annotators to train these models. To determine the final label for each instance, we combined the outputs of each cohort using various methods. Some methods were based on the labels themselves, while others utilized the probability distributions returned by the models:

- **Majority vote:** This approach determines the label by considering the majority vote among the six models. In case of a tie, we decided to set the label as "NO" in the first task. For the second task, the order of preference for tie-breaking is JUDGMENTAL > REPORTED > DIRECT.
- **Gender vote:** This approach adds, for each instance, the number of "votes" that each label receives from each gender's models. The label that receives more votes is selected. In case of a tie (two different labels obtain the same number of votes from the two genders' models), then the label selected by the "females" is returned as it has shown better results, as indicated in table 3.
- **Age vote:** This approach involves incorporating the "votes" received by each label from the models of each age group (18-22, 23-45, and 46 or more) for each instance. In each age group, we adopt the same decision rule as the previous method. In the case of a tie, the preferred label for the first task is "NO". For the second task, the order of preference is JUDGMENTAL > REPORTED > DIRECT.
- **Probability distribution mean:** This method computes the mean probability for each label and the six models, and the label with the highest mean probability is returned.
- **Probability distribution gender:** This method computes the most probable label for each gender by adding the scores of each gender's model. The label with the highest probability is selected. Since it is based on probability, it is assumed to be very difficult to have a tie-breaking situation.
- **Probability distribution Age:** This method computes the most probable label for each age range by adding the scores of each range's model. The label with the highest probability is selected.
- **Probability distribution cohort:** This method calculates the probability for each label/-cohort by considering the probability outputted by the model. The label with the highest probability is then selected and returned as the final prediction.

Finally, we employed all of these methods to train models using the augmented version of the EXIST 2023 train set and assessed their performance on the EXIST 2023 development set. Table 3 presents the results for Task 1.

Based on the obtained results, we have selected the outputs from the "Learning from raw disagreements" method as the first run for Task 1. This method has exhibited promising performance, and we will also employ its results to filter out instances labeled as NO-SEXIST, as it has demonstrated high accuracy in classifying this label.

As the second run for Task 1, we have submitted the "Gender vote" method. This approach consolidates the votes of the annotators from each gender to determine the final label for each instance.

For the third run in Task 1, we employ the outputs from the "Probability distribution (age)" method. This approach considers the probability distributions assigned by each age range cohort to determine the final label for each instance.

Table 3
TASK 1 evaluation results on the EXIST development dataset

Method	ICM hard-hard	F-measure hard-hard (YES)	F-Measure hard-hard (NO)	F-measure hard-hard (macro-F)	ICM hard-soft	ICM soft-soft
Learning from raw disagreements	0.5873	0.8068	0.8272	0.817	0.4393	0.7447
Majority vote	0.5841	0.8059	0.826	0.816	0.4563	-
Age vote	0.573	0.8	0.8246	0.8123	0.4214	-
Gender vote Male preference	0.5883	0.8091	0.8255	0.8173	0.4458	0.3913
Gender vote Female preference	**0.5942**	**0.8105**	0.8279	**0.8192**	**0.4758**	0.4125
Probability distribution	0.575	0.8029	0.8233	0.8131	0.4306	-
Probability distribution (mean)	0.575	0.8029	0.8233	0.8131	0.4306	0.7363
Probability distribution (age)	0.5785	0.8037	0.8247	0.8142	0.4274	-1.5735
Probability distribution (gender)	0.575	0.8029	0.8233	0.8131	0.4306	0.785
Probability distribution (cohort)	0.5841	0.8038	**0.828**	0.8159	0.4485	**0.8499**

A similar approach is followed for addressing Task 2. This is a multi-class classification Task instead of a binary classification one. Initially, tweets labeled as non-sexist (from Task 1) were eliminated using the approach that achieved the best F-Measure score for the "NO" (non-sexist) class, which involved using the Probability distribution cohort method. Subsequently, we repeated the previous steps to predict the source intention. The results of this approach evaluated on the development dataset are displayed in Table 4.

Table 4
TASK 2 evaluation results on the EXIST development dataset

Method	ICM hard-hard	FMeasure hard-hard (JUDGEMENTAL)	F-Measure hard-hard (NO)	FMeasure hard-hard (REPORTED)	F-measure hard-hard (DIRECT)	F-measure hard-hard (macro-F)	ICM hard-soft	ICM soft-soft
Learning from raw disagreement	0.2936	**0.3178**	0.828	0.3394	0.5945	0.5199	-6.8743	-
Majority vote	0.2936	**0.3178**	0.828	0.3394	0.5945	0.5199	-6.8743	-
Gender vote	0.2811	0.2812	0.828	0.3293	0.5951	0.5084	-6.6572	-
Age vote	0.3106	0.3015	0.828	0.3509	**0.6083**	0.5222	-6.7823	-
Probability distribution	0.29	0.2667	0.828	**0.3681**	0.5935	0.5141	-6.6209	-
Probability distribution (mean)	**0.3142**	0.2974	0.828	0.3659	0.5988	**0.5225**	-6.5681	-1.2093
Probability distribution (age)	0.279	0.2588	0.828	0.3294	0.5922	0.5021	**-6.4868**	-2.2366
Probability distribution (gender)	0.3021	0.3099	0.828	0.358	0.5937	0.5224	-6.8162	**-1.1792**

We have selected the outputs from the "Majority vote" method as the first run for Task 2, the "Probability distribution (mean)" method for the second run, and the "Probability distribution (gender)" method for the third run.

5. Evaluation and results

Table 5 provides a summary of the strategies employed in each of the runs that were ultimately submitted for both tasks.

For each of the tasks, the organization performed three types of evaluations:

- Hard-hard: the hard system output is compared against the hard ground truth.
- Hard-soft: the hard system output is compared against the soft ground truth.
- Soft-soft: the soft system output is compared against the soft ground truth.

For all tasks and evaluation types (hard-hard, hard-soft, and soft-soft), the official metric used is ICM (Information Contrast Measure) [24]. ICM is a similarity function that extends Pointwise

Table 5
Strategies employed in the runs submitted for Task 1 and Task 2

Task	Run	Method
Task 1	1	Repeated labeling
Task 1	2	Gender vote
Task 1	3	Probability distribution (cohorts)
Task 2	1	Majority vote
Task 2	2	Probability distribution (mean)
Task 2	3	Probability distribution (gender)

Mutual Information (PMI) and is employed to evaluate system outputs in classification problems by measuring their similarity to the ground truth categories. An extended version of ICM, known as ICM-soft, has been specifically developed for the task to accommodate both soft system outputs and soft ground truth assignments. The results of our three runs for Task 1 and Task 2, evaluated using the three types of evaluation, are presented in Tables 6, 7, 8, 9, 10, and 11. Each table provides details on various evaluation metrics for both Spanish and English languages, as well as the combined results for both languages. In each table, the first column within each column group indicates the ranking position of each run. The first and second rows of values in each table represent the gold and best results achieved for the respective task in each evaluation.

5.1. Task 1 - Sexism Identification

Regarding Task 1, as shown in Tables 6, 7, and 8, the third run utilizing the 'probability distribution (cohorts)' method outperformed the others in terms of hard metrics for all languages. However, the first run, based on 'repeated labeling,' achieved better results in the soft-soft evaluation. Our best approach secured the 19th position out of 57 participants in the hard-hard evaluation, the 17th position in the hard-soft evaluation, and the 12th position in the soft-soft evaluation. The higher ranking in the soft evaluations suggests that our approach effectively captures the different perceptions of sexism among distinct population cohorts.

Table 6
Results for the hard-hard evaluation for Task 1: This table presents the results obtained in the hard-hard evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM, Normalized ICM, and F-measure. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL				ES				EN			
	Rank	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1
Gold	0	0.9948	1	1		0.9999	1	1		0.9798	1	1
Best score	1	0.6575	0.785	0.8109	1	0.6995	0.8011	0.8384	1	0.6004	0.7693	0.776
JPM_UNED_1	28	0.5057	0.6883	0.756	20	0.514	0.6783	0.7748	30	0.4819	0.6972	0.7308
JPM_UNED_2	33	0.4863	0.6759	0.7533	24	0.5016	0.6701	0.7784	39	0.4556	0.6812	0.7204
JPM_UNED_3	19	**0.5223**	**0.6989**	**0.7623**	14	**0.545**	**0.6988**	**0.7885**	29	**0.4844**	**0.6987**	**0.7284**

Table 7
Results for the hard-soft evaluation for Task 1: This table presents the results obtained in the hard-soft evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM-Soft and ICM-Soft normalized. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL			ES			EN		
	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm
Gold	0	3.1182	1	0	3.1177	1	0	3.1141	1
Best score	1	1.1977	0.6897	1	1.3487	0.6892	1	0.9695	0.6905
JPM_UNED_1	19	0.1685	0.5235	17	0.3485	0.5135	31	-0.1243	0.5327
JPM_UNED_2	32	0.1075	0.5136	23	0.2927	0.5037	38	-0.1879	0.5235
JPM_UNED_3	17	**0.2041**	**0.5292**	13	**0.3927**	**0.5212**	28	**-0.0924**	**0.5373**

Table 8
Results for the soft-soft evaluation for Task 1: This table presents the results obtained in the soft-soft evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM-Soft, ICM-Soft normalized and Cross entropy. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL				ES				EN			
	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Gold	0	3.1182	1	0.5472	0	3.1177	1	0.5208	0	3.1141	1	0.577
Best score	1	0.903	0.6421	0.796	1	0.9527	0.6196	0.7672	1	0.8157	0.6683	1.0198
JPM_UNED_1	12	**0.6779**	**0.6058**	**0.8023**	18	0.6536	0.5671	**0.7588**	11	**0.6632**	**0.6463**	0.8512
JPM_UNED_2	18	0.5972	0.5927	**0.8852**	16	**0.6641**	**0.5689**	0.8116	15	0.4853	0.6207	0.9677
JPM_UNED_3	29	0.2467	0.5361	2.2342	31	0.1576	0.4799	2.1581	24	0.3506	0.6012	2.3196

5.2. Task 2 - Source Intention Identification

For the second task (categorizing the tweets according to the intention of the source), as shown in Tables 9, 10 and 11, the second run utilizing the 'probability distribution (mean)' method obtains the best results in this task and achieves the second position in hard-soft evaluation for Spanish. Over both languages, our best approach secured the 9th position in the hard-hard evaluation, the 22nd position in the hard-soft evaluation, and the 3rd position in the soft-soft evaluation.

Table 9
Results for the hard-hard evaluation for Task 2: This table presents the results obtained in the hard-hard evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM, Normalized ICM, and F-measure. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL				ES				EN			
	Rank	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1
Gold	0	1.5378	1	1	0	1.6007	1	1	0	1.4449	1	1
Best score	1	0.4887	0.7764	0.5715	1	0.5711	0.7732	0.6059	1	0.3677	0.781	0.5224
JPM_UNED_1	11	0.1673	0.7079	0.5032	10	0.1986	0.6911	0.5281	12	0.1024	0.727	0.4661
JPM_UNED_2	9	**0.1862**	**0.712**	0.5054	8	**0.2351**	**0.6992**	0.5341	13	0.0995	0.7264	0.4649
JPM_UNED_3	10	0.1806	0.7108	**0.5092**	9	0.2231	0.6965	**0.5383**	11	**0.1034**	**0.7272**	**0.4673**

Table 10
Results for the hard-soft evaluation for Task 2: This table presents the results obtained in the hard-soft evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM-Soft and ICM-Soft normalized. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL			ES			EN		
	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm
Gold	0	6.2057	1	0	6.2431	1	0	6.1178	1
Best score	1	-2.3974	0.7803	1	-1.8502	0.7684	1	-3.265	0.7943
JPM_UNED_1	23	-7.5078	0.6498	21	-6.8073	0.6266	25	-8.9034	0.6707
JPM_UNED_2	22	**-7.3346**	**0.6542**	19	**-6.5622**	**0.6336**	23	**-8.8583**	**0.6717**
JPM_UNED_3	24	-7.5205	0.6495	22	-6.8533	0.6253	24	-8.8978	0.6708

Table 11
Results for the soft-soft evaluation for Task 2: This table presents the results obtained in the soft-soft evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM-Soft, ICM-Soft normalized and Cross entropy. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL				ES				EN			
	Rank	ICM-Soft	ICM-Hard Soft	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Gold	0	6.2057	1	0.9128	0	6.2431	1	0.8926	0	6.1178	1	0.9354
Best score	1	-1.3443	0.8072	1.7833	1	-1.2317	0.7861	1.6415	1	-1.1471	0.8407	1.8001
JPM_UNED_2	3	**-1.675**	**0.7988**	**2.5549**	2	**-1.4414**	**0.7801**	**2.4472**	6	-2.1062	0.8197	2.6757
JPM_UNED_3	5	-1.6888	0.7984	2.5561	3	-1.5006	0.7785	2.4511	4	**-2.0436**	**0.8211**	**2.674**

6. Conclusions

This paper presents the participation of the JPM-UNED team in the Task 1 and the Task 2 of the EXIST 2023 Lab at CLEF, which focuses on the classification of sexism in social networks with disagreement. We have investigated different approaches to learning with disagreement, leveraging the current state of the art. Furthermore, we essayed different data augmentation techniques, such as incorporating translations of tweets between English and Spanish in the training set.

Among the approaches we explored, some of them proposed different variations of the judgement aggregation method, which combines the judgments or opinions of the multiple annotators and models to arrive at the final label. Notably, our best results were obtained in Task 2, where we secured the second position in the soft-soft evaluation for the Spanish language. This achievement was made possible by employing the Judgement Aggregation approach that leverages the viewpoints of the six different cohorts.

One limitation of our work stems from the size of the dataset. As some of our approaches involved splitting the dataset into six cohorts, the resulting training datasets were relatively small, which presented challenges in effectively training the models.

7. Acknowledgments

This research is funded by FAIRTRANSNLP-DIAGNOSIS: Measuring and quantifying bias and fairness in NLP systems, grant PID2021-124361OB-C32, funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. This work has been also funded by the Ministry of Universities and the European Union through the EuropeaNextGenerationUE funds and the "Plan de Recuperación, Transformación y Resiliencia".

References

[1] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[2] A. R. G. Prasad, V. Seshagiri, L. Ravindranath, et al., A catalytic spectrophotometric method for the analytical determination of trace amounts of mercury (ii), Chem. Sci. Jour (2010) 1–8.

[3] S. Y. Jeong, Y. S. Koh, G. Dobbie, Phishing detection on twitter streams, in: Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2016 Workshops, BDM, MLSDA, PACC, WDMBF, Revised Selected Papers 20, Springer, 2016, pp. 141–153.

[4] J. Nogués Graell, Detection of toxicity in social media. a study on semantic orientation and linguistic structure (2022).

[5] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[6] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[7] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023, pp. 593–599.

[8] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Semeval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 338–347.

[9] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 – learning with disagreement for sexism identification and characterization(extended overview), in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[10] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 – learning with disagreement for sexism identification and characterization, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[11] O. Istaiteh, R. Al-Omoush, S. Tedmori, Racist and sexist hate speech detection: Literature

review, in: 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), IEEE, 2020, pp. 95–99.
[12] H. Abburi, S. Sehgal, H. Maheshwari, V. Varma, Knowledge-based neural framework for sexism detection and classification., in: Proceedings of IberLEF@ SEPLN, 2021, pp. 402–414.
[13] L. Altin, H. Saggion, Automatic detection of sexism in social media with a multilingual approach, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings Series, 2021, pp. 415–419.
[14] H. Abburi, P. Parikh, N. Chhaya, V. Varma, Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach, Data Sci. Eng. 6 (2021) 359–379.
[15] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, volume 2253, CEUR-WS, 2018, pp. 1–6.
[16] L. I. Merlo, B. Chulvi, R. Ortega, P. Rosso, When humour hurts: linguistic features to foster explainability, 2023-03.
[17] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm, J. Royal Stat. Soc. Series 28 (1979) 20–28.
[18] F. Rodrigues, F. Pereira, B. Ribeiro, Gaussian process classification and active learning with multiple annotators, in: Proceedings of ICML, 2014, pp. 433–441.
[19] P. Welinder, P. Perona, Online crowdsourcing: Rating annotators and obtaining cost-effective labels, in: IEEE CVPR Workshops, 2010, pp. 25–32.
[20] D. Reidsma, R. Akker, Exploiting 'subjective' annotations, in: Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, 2008, pp. 8–16.
[21] V. S. Sheng, F. Provost, P. G. Ipeirotis, Get another label? improving data quality and data mining using multiple, noisy labelers, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 614–622.
[22] F. Rodrigues, M. Lourenço, B. Ribeiro, F. C. Pereira, Learning supervised topic models for classification and regression from crowds, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 2409–2422.
[23] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.
[24] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 2022, pp. 5809–5819.