

La investigación reportada en esta tesis es parte de los programas de investigación del CICESE (Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California).

La investigación fue financiada por el SECIHTI (Secretaría de Ciencia, Humanidades, Tecnología e Innovación).

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México). El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo o titular de los Derechos de Autor.

CICESE © 2025, Todos los Derechos Reservados, CICESE

Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California



Maestría en Ciencias en Ciencias de la Computación

Identificación de sexismo en redes sociales

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de
Maestra en Ciencias

Presenta:

Martha Paola Jimenez Martinez

Ensenada, Baja California, México

2025

Tesis defendida por

Martha Paola Jimenez Martinez

y aprobada por el siguiente Comité

Dr. Irvin Hussein López Nava

Codirector de tesis

Dr. Manuel Montes y Gómez

Codirector de tesis

Dr. Jesús Favela Vara

Dra. Rufina Hernández Martínez



Dr. Pedro Gilberto López Mariscal

Coordinador del Posgrado en Ciencias de la Computación

Dra. Ana Denise Re Araujo

Directora de Estudios de Posgrado

Resumen de la tesis que presenta Martha Paola Jimenez Martinez como requisito parcial para la obtención del grado de Maestra en Ciencias en Ciencias de la Computación.

Identificación de sexismo en redes sociales

Resumen aprobado por:

Dr. Irvin Hussein López Nava

Codirector de tesis

Dr. Manuel Montes y Gómez

Codirector de tesis

La presente investigación aborda la detección de sexismo en redes sociales en español, empleando técnicas de procesamiento de lenguaje natural y modelos avanzados basados en transformers. Ante la escasez de recursos en español, se utilizó el corpus EXIST 2023, anotado por diversos perfiles demográficos, para analizar cómo el género y la edad influyen en la percepción y clasificación del sexismo mediante tres niveles de tareas: identificación de sexismo (clasificación binaria para determinar si un tuit es sexista), intención de la fuente (clasificación de mensajes sexistas según intención: directa, reportada o juicio), y categorización del tipo de sexismo (agrupación en categorías como ideología, estereotipos, cosificación, violencia sexual y misoginia). A través de una metodología que incluye el preprocesamiento de datos, la clasificación en tareas binarias, tanto multiclase y como multietiqueta, y el desarrollo de ensambles que incorporan perspectivas demográficas, se logró un puntaje F1 de 0.854. Además, se realizó un análisis de errores para identificar las limitaciones del modelo y se implementó ChatGPT para generar explicaciones contextuales de las predicciones, mejorando la interpretabilidad del sistema. Los hallazgos subrayan la relevancia de integrar enfoques inclusivos y técnicas de explicabilidad para abordar un problema culturalmente complejo como el sexismo en entornos digitales. Este trabajo no solo contribuye al avance en la detección de lenguaje discriminatorio en español, sino que también establece bases para desarrollar sistemas más justos y transparentes en tareas críticas de análisis textual.

Palabras clave: Sexismo, Redes de Atención Jerárquica, Transformers, Clasificación por Ensamble, Detección de Sexismo, Procesamiento de Lenguaje Natural (PLN), Explicabilidad, ChatGPT, Análisis Demográfico, Clasificación Multietiqueta

Abstract of the thesis presented by Martha Paola Jimenez Martinez as a partial requirement to obtain the Master of Science degree in .

Identification of Sexism on Social Media

Abstract approved by:

Dr. Irvin Hussein López Nava

Thesis Co-Director

Dr. Manuel Montes y Gómez

Thesis Co-Director

The present research addresses the detection of sexism on Spanish-language social media using natural language processing techniques and advanced transformer-based models. Given the scarcity of resources in Spanish, the EXIST 2023 corpus, annotated by diverse demographic profiles, was utilized to analyze how gender and age influence the perception and classification of sexism across three task levels: Sexism identification (binary classification to determine if a tweet is sexist), source intent (classification of sexist messages based on intent: direct, reported, or judgmental), and categorization of sexism type (grouping into categories such as ideology, stereotypes, objectification, sexual violence, and misogyny). Through a methodology encompassing data preprocessing, binary, multiclass, and multilabel classification tasks, and the development of ensembles incorporating demographic perspectives, the study achieved an F1 score of 0.854. Additionally, an error analysis was conducted to identify the model's limitations, and ChatGPT was implemented to generate contextual explanations of predictions, enhancing the system's interpretability. The findings highlight the importance of integrating inclusive approaches and explainability techniques to address a culturally complex issue like sexism in digital environments. This work not only contributes to advancements in the detection of discriminatory language in Spanish but also establishes foundations for developing fairer and more transparent systems in critical text analysis tasks.

Keywords: Sexism, Hierarchical Attention Networks, Transformers, Ensemble Classification, Sexism Detection, Natural Language Processing (NLP), Explainability, ChatGPT, Demographic Analysis, Multilabel Classification

Dedicatoria

A mis queridos padres, por su apoyo incondicional. Y a todas las personas que, con su ayuda y aliento, me acompañaron en este camino.

Agradecimientos

Al Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California, por brindarme la oportunidad de formar parte de su programa de posgrado.

A la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), por el invaluable apoyo económico.

A mis directores de tesis, el Dr. Irvin Hussein López Nava, por guiarme hacia un tema que realmente me apasiona y por todas las valiosas enseñanzas compartidas; y al Dr. Manuel Montes y Gómez, por su confianza en mí, su paciencia y su constante apoyo.

A mi comité de tesis, el Dr. Jesús Favela Vara y la Dra. Rufina Hernández Martínez, por su interés en formar parte de este trabajo, sus recomendaciones enriquecedoras y su constante ánimo.

Tabla de contenido

	Página
Resumen en español	ii
Resumen en inglés	iii
Dedicatoria	iv
Agradecimientos	v
Lista de figuras	viii
Lista de tablas	ix
Capítulo 1. Introducción	
1.1. Planteamiento del problema	3
1.2. Pregunta de investigación	4
1.3. Objetivos	4
1.3.1. Objetivo general	4
1.3.2. Objetivos específicos	4
1.4. Metodología	5
1.5. Estructura de la tesis	6
Capítulo 2. Marco Teórico	
2.1. Clasificación de texto	7
2.1.1. Transformers	7
2.1.2. Modelos basados en transformers	9
2.1.2.1. BERT	10
2.1.2.2. RoBERTuito	10
2.2. Explicabilidad	11
2.2.1. Valores de Atención	12
2.2.2. Explicabilidad usando redes de atención jerárquicas	12
2.2.3. ChatGPT	13
2.3. Resumen del capítulo	14
Capítulo 3. Trabajo relacionado	
3.1. Competencia EXIST: Ediciones 2023 y 2024	16
3.1.1. Identificación de sexismo	17
3.1.2. Intención de la fuente	18
3.1.3. Categorización de tipo de sexismo	18
3.1.4. Métricas de evaluación	19
3.2. Revisión de la literatura	20
3.2.1. Principales técnicas y estrategias empleadas	20
3.2.1.1. Preprocesamiento y preparación de datos	20
3.2.1.2. Modelos basados en transformers	21
3.2.1.3. Uso de Modelos de Lenguaje Grandes para explicabilidad	22
3.2.1.4. Incorporación de información de anotadores	22
3.3. Comparación de desempeño y enfoques en las competencias EXIST 2023 y 2024 ..	23

Capítulo 4. Análisis de Percepciones sobre el Sexismo	
4.1. Conjunto de datos EXIST	27
4.1.1. Muestreo de datos	29
4.2. Análisis de las diferencias en la percepción del sexismo	30
4.2.1. Percepción del sexismo por género	31
4.2.2. Percepción del sexismo por edad	33
4.3. Resumen del capítulo	35
Capítulo 5. Detección Automática de Sexismo mediante Ensamblés	
5.1. Preprocesamiento de Datos y Metodología de Clasificación	36
5.2. Enfoque Propuesto Basado en Ensamble	37
5.3. Tarea 2: Intención	40
5.4. Tarea 3: Categorización	41
5.5. Explicación de las predicciones del modelo con ChatGPT	42
5.6. Resumen del capítulo	45
Capítulo 6. Resultados	
6.1. Preprocesamiento de los datos	46
6.2. Identificación Automática del Sexismo	47
6.3. Análisis de Error	55
6.4. Explicaciones de los Tuits Clasificados con ChatGPT	59
6.5. Resumen del capítulo	62
Capítulo 7. Conclusiones	
7.1. Discusión	64
7.2. Limitaciones	65
7.3. Contribuciones	66
7.4. Artículos derivados de esta investigación	67
7.5. Trabajo a Futuro	68
Literatura citada	70
Anexos	73

Lista de figuras

Figura	Página
1. Estructura global del codificador-decodificador.	8
2. Funcionamiento de ChatGPT.	14
3. Modelo general para las tres tareas.	38
4. Modelos para la Identificación de Sexismo	39
5. Modelo para la Detección de Intención.	41
6. Modelo para la Categorización del Sexismo.	42
7. Prompt utilizado.	44
8. Comparación del rendimiento del modelo al predecir con diferentes tipos de datos.	47
9. Identificación del Sexismo (Tarea 1).	49
10. Detección de la Intención (Tarea 2).	52
11. Categorización del Sexismo (Tarea 3).	54
12. Nube de Palabras Representando los Términos Comunes en los Tuits Clasificados como Falsos Positivos en el Modelo de Ensamble para Detectar Sexismo.	57
13. Nube de Palabras Representando los Términos Comunes en los Tuits Clasificados como Falsos Negativos en el Modelo de Ensamble para Detectar Sexismo.	58
14. Cruces por edad.	75
15. Cruces de hombres por distintas edades.	75
16. Cruces de mujeres por distintas edades.	76

Lista de tablas

Tabla	Página
1. Resultados de evaluación para las tareas de clasificación en español: detección de discurso de odio, análisis de sentimientos, análisis de emociones y detección de ironía. Los resultados se expresan como la media de la puntuación Macro F1 de 10 ejecuciones de los experimentos de clasificación (Pérez et al. (2021)).	11
2. Tabla Comparativa EXIST 2023 y 2024	23
3. Métodos y resultados de los equipos participantes en las tareas del EXIST 2023 (Plaza et al., 2023) y 2024 (Plaza et al., 2024).	25
4. Estadísticas de etiquetas duras del conjunto de datos EXIST 2023 (Plaza et al., 2023).	30
5. Matriz de Confusión de coincidencias y desacuerdos en las etiquetas de sexismo entre Hombres y Mujeres.	32
6. Coincidencias y diferencias en el etiquetado según género	32
7. Diferencias y similitudes entre personas de 18 a 22 años, de 23 a 45 años y de 46 años o más.	33
8. Coincidencia y desacuerdo en las etiquetas de sexismo entre diferentes grupos de edad.	34
9. Matriz de Confusión de Predicciones.	56

Capítulo 1. Introducción

En la era digital, las redes sociales se han convertido en una parte integral de nuestra vida cotidiana, facilitando la comunicación y la conexión entre personas de todo el mundo. Sin embargo, este fenómeno también ha traído consigo un aumento significativo en la propagación del discurso de odio en línea. Esta problemática ha generado una creciente preocupación entre organizaciones de derechos civiles, que exigen una aplicación más estricta de políticas para controlar y reducir este tipo de contenido. En respuesta, en los últimos años se han desarrollado múltiples métodos basados en el Procesamiento de Lenguaje Natural (PLN) para detectar automáticamente expresiones ofensivas o abusivas en estas plataformas (Ocampo et al., 2023). A pesar de los avances significativos en el uso de modelos de aprendizaje automático y técnicas de aprendizaje profundo para abordar este problema, la detección automática del discurso de odio sigue enfrentando importantes desafíos. La ambigüedad del lenguaje, la variabilidad cultural en las definiciones de discurso de odio, y la falta de datos estandarizados en múltiples idiomas complican el desarrollo de soluciones robustas. Además, los estudios recientes han destacado que, aunque los modelos basados en BERT y otras arquitecturas de aprendizaje profundo han mejorado el rendimiento, aún persisten limitaciones en la generalización y precisión de los modelos para identificar de manera efectiva diferentes tipos de discurso ofensivo, como el abuso, el ciberacoso y la radicalización. Esto resalta la necesidad de investigaciones continuas y la creación de conjuntos de datos multilingües más completos para superar estas barreras (Jahan & Oussalah, 2023).

Este fenómeno ha despertado un creciente interés académico, evidenciado por el aumento en el número de publicaciones sobre discurso de odio indexadas en la Web of Science (WoS), que pasaron de 42 en 2013 a 162 en 2018. Este incremento no solo refleja la prioridad otorgada a la investigación sobre este fenómeno, sino que también subraya el impacto del discurso de odio en las sociedades contemporáneas, donde su presencia en los medios de comunicación y en internet es cada vez más notoria. La investigación sobre este tema abarca múltiples disciplinas, incluyendo derecho, sociología, comunicación y psicología, lo que resalta su complejidad y relevancia social (Paz et al., 2020).

El término “discurso de odio” no solo describe un tipo específico de expresiones, sino que también opera como una herramienta evaluativa, emitiendo un juicio negativo sobre su contenido y considerándolo merecedor de censura (Anderson & Barnes, 2022). Este concepto es especialmente relevante al abordar el sexismo en línea, un tipo de discurso de odio que perpetúa la discriminación de las personas por razón de sexo (Real Academia Española, 2024).

Así, el sexismo en línea no solo refuerza estereotipos que benefician a un género sobre el otro, sino

que también fomenta la idea de la inferioridad de la mujer debido a sus diferencias biológicas (Comisión Nacional para Prevenir y Erradicar la Violencia Contra las Mujeres, 2016). Esta percepción se ha arraigado profundamente en diversas culturas a lo largo de la historia. Incluso en el Código Civil Francés, que influyó significativamente en las primeras leyes mexicanas, se consideraba a las mujeres como dependientes de los hombres en todos los aspectos de la vida, destinándolas al ámbito doméstico mientras que los hombres eran vistos como seres independientes y capaces de pensar y actuar por su cuenta (de la Morena, Ines, 2020).

El lenguaje es una de las maneras más sutiles de transmitir discriminación, ya que refleja los valores y pensamientos de la sociedad que lo crea y utiliza. Aunque algunas palabras pueden parecer neutras en su uso cotidiano, muchas otras están cargadas de significados que refuerzan estereotipos y roles de género (Secretaría de Salud, 2011).

Es muy común que cualquier persona nacida o residente en México, así como en otras culturas latinas, reconozca la fuerte influencia del machismo. Todos hemos escuchado al menos una de las muchas frases sexistas que son frecuentes en nuestro lenguaje (CONAVIM, 2018).

Un estudio sobre “Lenguaje Misógino en Twitter y Violencia Sexual” revela que el uso de lenguaje misógino en las redes sociales está correlacionado con un mayor número de violaciones a nivel estatal en Estados Unidos. El seguimiento del lenguaje en las redes sociales y otros espacios públicos en Internet podría, por lo tanto, proporcionar una fuente alternativa de información sobre el nivel de violencia sexual en la población (Fulper et al., 2014).

En el contexto del sexismo, se transmiten los estereotipos y roles históricamente asignados a mujeres y hombres en una sociedad (CONAVIM, 2016). Las expresiones misóginas, como las utilizadas en redes sociales, buscan ejercer presión utilizando la figura femenina como sinónimo de debilidad u ofensa, reforzando la idea de que los valores masculinos son el estándar de lo deseable y bueno. Esto implica que ser mujer, o tener atributos femeninos, es considerado inferior (CONAVIM, 2018).

Por otro lado y también en un mismo entorno machista, los hombres también pueden ser víctimas, ya que existen expresiones destinadas obligarlos a ser fuertes, exitosos, a reprimir sus emociones o a comportarse de maneras con las que pueden no sentirse cómodos. Ejemplos de estas frases son “los hombres no lloran” o “no seas nena” (CONAVIM, 2018).

1.1. Planteamiento del problema

Richards et al. (2015) revela que el efecto de las redes sociales en la salud de los niños y jóvenes es especialmente significativo en su salud mental, afectando en particular su autoestima y bienestar, y está asociado con problemas como el ciberacoso. A partir de la pandemia global de 2020, las redes sociales se han transformado en un recurso valioso, aunque también en una amenaza potencial. La gran cantidad de información publicada en redes sociales en un breve lapso de tiempo ha tenido un impacto notable (Venegas-Vera et al., 2020). Uno de los problemas emergentes más preocupantes es la proliferación de discursos de odio en estas plataformas, que afectan a diversas comunidades y grupos vulnerables. Dentro de esta problemática, destaca particularmente la presencia de lenguaje sexista, un fenómeno que requiere especial atención. La detección automática de este tipo de contenido enfrenta retos significativos debido a la complejidad del lenguaje, la sutileza de las expresiones discriminatorias y las variaciones culturales y lingüísticas.

A pesar de los resultados prometedores obtenidos por los modelos de aprendizaje profundo, como los transformers, estos aún no son capaces de proporcionar una explicabilidad clara de manera directa. La explicabilidad es crucial porque permite entender cómo y por qué un modelo toma ciertas decisiones, lo que resulta fundamental en tareas sensibles como la detección de lenguaje sexista. Sin una explicabilidad adecuada, es difícil evaluar si las decisiones del modelo están alineadas con criterios éticos, libres de sesgos y culturalmente apropiados, lo que podría limitar su aplicabilidad práctica y su aceptación por parte de los usuarios. Aunque ciertos mecanismos internos de los modelos transformers, como la atención, pueden sugerir transparencia al asignar pesos a las entradas, en la práctica, estos modelos suelen comportarse como “cajas negras”. Esto significa que, a pesar de que se puedan interpretar algunos de sus componentes, la relación real entre los procesos internos del modelo y sus decisiones finales sigue siendo difícil de entender (Jain & Wallace, 2019).

Por otro lado, la mayoría de los recursos disponibles se centran en inglés, mientras que otros idiomas, como el español, cuentan con una cantidad considerablemente menor de recursos.

Por lo tanto, esta tesis no solo pretende abordar el problema de la detección de sexismo en español mediante su identificación, clasificación y análisis de intención, sino que también explora la explicabilidad del modelo utilizado. El objetivo es ofrecer una comprensión más profunda del funcionamiento interno del modelo de clasificación, proporcionando transparencia sobre cómo se toman las decisiones y mejorando así la interpretabilidad de los resultados obtenidos.

1.2. Pregunta de investigación

- ¿Cómo varía la valoración del sexismo en textos en español según el género y la edad de los anotadores de redes sociales?
- ¿Es más efectivo utilizar un único modelo transformer generalista o un ensamble de modelos especializados en perfiles de usuarios (por género o edad) para la clasificación del sexismo en textos en español?
- ¿Cómo contribuyen los modelos basados en transformers al análisis explicativo en la detección de contenido sexista en textos en español?

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar y evaluar un método para la detección de sexismo en español que incorpore en su entrenamiento las valoraciones de personas de distintos géneros y rangos de edad, con el fin de mejorar la explicabilidad de los modelos del estado del arte.

1.3.2. Objetivos específicos

- Analizar la subjetividad en la percepción y valoración de comentarios sexistas expresados en redes sociales, evaluando cómo anotadores de distintos géneros y rangos de edad coinciden o discrepan en sus juicios, para entender mejor las diferencias intersubjetivas en la anotación.
- Diseñar y evaluar un método de clasificación de comentarios sexistas basado en un ensamble de modelos especializados en los distintos grupos de individuos (según género y edad), con el fin de mejorar la precisión en la detección y categorización del sexismo.
- Explorar el uso de grandes modelos de lenguaje para generar explicaciones argumentativas sobre las decisiones del clasificador, aumentando la interpretabilidad de los resultados sin sacrificar el

rendimiento en las tareas de identificación, distinción de intención y categorización del tipo de sexismo.

1.4. Metodología

La metodología de esta investigación se desarrolló a través de los siguientes pasos:

1. **Adquisición de una Colección de Tuits sobre Sexismo:** Se recopiló y analizó una base de datos relevante sobre sexismo en español, utilizando el conjunto de datos *EXIST 2023* (Plaza et al., 2023). Antes del análisis, se realizó un preprocesamiento de los datos para garantizar su adecuación a las tareas de clasificación.
2. **Análisis exploratorio de las etiquetas por perfil:** Se llevó a cabo un análisis de cómo fueron categorizados los tuits según el género y la edad de los anotadores, identificando posibles patrones o sesgos en función de estos perfiles.
3. **Diseño de Métodos de Clasificación Binaria, Multiclase y Multietiqueta con Información Textual:** A partir de la base de datos *EXIST 2023*, se implementaron y evaluaron modelos de clasificación en tres modalidades:
 - Clasificación binaria, para distinguir entre tuits sexistas y no sexistas.
 - Clasificación multiclase, para categorizar los tuits sexistas en subtipos específicos (directo, denunciado, juzgado).
 - Clasificación multietiqueta, para identificar la presencia de diversas formas de sexismo (ideológico, estereotipo, objetificación, violencia sexual y misoginia).
4. **Implementación de Ensamblados:** Se exploraron distintas técnicas de ensamble para mejorar la precisión y robustez de los modelos de clasificación. Aprovechando la categorización de los datos basada en los perfiles de los anotadores, se diseñaron ensambles específicos que integraron las características y perspectivas de estos perfiles.
5. **Análisis de Error:** Se realizó un análisis detallado de los errores cometidos por el modelo de ensamble, identificando las palabras y patrones textuales a los que el modelo prestó mayor atención durante la clasificación. Este análisis ayudó a comprender mejor las limitaciones del modelo y las áreas donde se podría mejorar.

6. **Explicabilidad de las Predicciones del Modelo:** Con el objetivo de aumentar la interpretabilidad de las predicciones, se utilizó *ChatGPT* para generar explicaciones detalladas de las predicciones. Estas explicaciones se basaron en el contenido y contexto específico de cada tuit, facilitando una comprensión más profunda de los resultados obtenidos.

1.5. Estructura de la tesis

- El **Capítulo 2** incluye una visión general de la información de contexto que es significativa para comprender el contenido de esta tesis. Revisamos los modelos tradicionales transformers y las redes de atención jerárquicas.
- En el **Capítulo 3** se describe trabajo previo relacionado con el problema de detección de comentarios sexistas en redes sociales, centrándonos principalmente en el análisis de trabajos del foro EXIST (Identificación de sexismo en redes sociales), que es una serie de eventos científicos y tareas compartidas sobre la identificación del sexismo en redes sociales.
- En el **Capítulo 4** se presenta nuestro análisis de la diferencia de la valoración de sexismo entre distintos anotadores, usando el corpus de EXIST.
- En el **Capítulo 5** se presenta la metodología propuesta basado en ensamble de modelos basados en transformers, ajustados cada uno a las valoraciones de un grupo de individuos particular.
- El **Capítulo 6** se centra en la evaluación del enfoque de transformers propuesto, examinando la integración de las Redes de Atención Jerárquica. Este enfoque busca mejorar la identificación y el análisis del sexismo en redes sociales.
- El **Capítulo 7** integra nuestras conclusiones, donde presentamos los principales hallazgos y limitaciones de esta tesis y sugerimos direcciones para futuras investigaciones en el campo del sexismo.

Capítulo 2. Marco Teórico

En este capítulo se presentan los antecedentes necesarios para comprender el enfoque y los métodos empleados en esta investigación. Se abordan los fundamentos de la clasificación de texto, con énfasis en los avances recientes basados en modelos preentrenados y arquitecturas como los transformers, así como las implicaciones de la explicabilidad en el procesamiento del lenguaje natural. Este capítulo no solo contextualiza la problemática, sino que también fundamenta las decisiones técnicas y metodológicas desarrolladas en los capítulos siguientes.

2.1. Clasificación de texto

La tarea de clasificación de texto es el proceso de asignar una etiqueta o categoría a un fragmento de texto en función de su contenido. Tradicionalmente, se ha utilizado un enfoque supervisado que combina el modelo Bag of Words (BoW) con un clasificador. En este enfoque, el texto se convierte en una representación basada en la frecuencia de aparición de palabras sin tener en cuenta el orden o contexto de las mismas, y luego se aplica un clasificador (como una regresión logística o máquinas de soporte vectorial) para hacer predicciones (Patil et al., 2023).

Sin embargo, enfoques recientes han optado por utilizar modelos pre-entrenados como *Word2Vec*, *GloVe* o *Bidirectional Encoder Representations from Transformers (BERT)*, que capturan el significado y el contexto de las palabras. Estos modelos permiten ajustar finamente la tarea específica de clasificación de texto, logrando mejores resultados.

Este trabajo utiliza este enfoque moderno, y en las siguientes secciones se explicarán los elementos principales de estos modelos pre-entrenados y su ajuste fino.

2.1.1. Transformers

Los transformers son una arquitectura de redes neuronales que revolucionó el procesamiento del lenguaje natural (PLN) y otras tareas de modelado secuencial. Propuestos en 2017 por investigadores de Google en el artículo "*Attention is all you need*", introdujeron una nueva estructura para abordar problemas relacionados con secuencias, como la traducción automática (Vaswani et al., 2017). A diferencia de las

arquitecturas recurrentes como las redes *Long Short-Term Memory (LSTM)*, que procesaban los datos de forma secuencial y dependían de un estado oculto para retener la información, los transformers emplean una técnica de atención que permite el procesamiento paralelo de las secuencias, mejorando tanto la eficiencia como la capacidad de modelado.

El uso de mecanismos de atención fue un avance clave, ya que en modelos anteriores, como las redes recurrentes, la información se comprimía en un solo estado oculto, lo que provocaba la pérdida de contexto, especialmente en secuencias largas. En cambio, los transformers introducen un mecanismo de atención llamado *Scaled Dot-Product Attention*, que calcula la relevancia entre cada palabra de la secuencia mediante el producto escalar entre las *queries* y las *keys*, escalado por la raíz cuadrada de la dimensión de las *keys*. Este enfoque permite que cada palabra se compare no solo consigo misma, sino con todas las demás palabras al mismo tiempo, lo que acelera el procesamiento de las secuencias Vaswani et al. (2017).

Además de su eficiencia en el procesamiento paralelo de secuencias, los transformers fueron originalmente diseñados para tareas de transducción de secuencias, como la traducción automática. Destacan por su capacidad para convertir secuencias de entrada en secuencias de salida mediante la auto-atención, sin la necesidad de utilizar redes neuronales recurrentes (RNNs) o convolucionales alineadas con la secuencia. La característica principal de los transformers es que mantienen la estructura de modelo encoder-decoder. El encoder toma la secuencia de entrada y genera una representación matricial de esta. Por ejemplo, la frase en inglés "How are you?" se convertiría en dicha representación interna. Luego, el decoder utiliza esa representación codificada para generar iterativamente una secuencia de salida, que en este caso sería la traducción en español: "¿Cómo estás?". Tanto el encoder como el decoder están compuestos por múltiples capas apiladas. Todos los encoders tienen la misma estructura y procesan la entrada secuencialmente, mientras que los decoders reciben tanto la información del último encoder como del decoder anterior en el proceso de generación de la salida (Ferrer, J., 2024) como se puede ver en la figura 1.

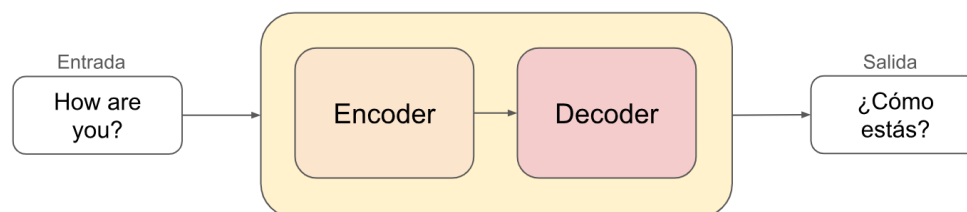


Figura 1. Estructura global del codificador-decodificador.

Cada codificador del transformer incluye dos subcapas: la capa de autoatención y la capa de prealimentación. La capa de autoatención permite al modelo evaluar la importancia de las palabras en función del contexto. Para ello, las palabras se dividen en vectores de consulta, clave y valor, y se aplica una función softmax para ponderar cada palabra, de manera que el modelo se centre en las más relevantes y minimice las irrelevantes. La salida de esta capa se procesa en la capa de prealimentación, una red neuronal aplicada en paralelo a cada posición de la secuencia, lo cual mejora la eficiencia computacional (Reddy, S., 2023).

Por otro lado, el decodificador contiene una estructura similar, con una capa adicional de atención de codificador-decodificador que permite enfocarse en las palabras importantes de la secuencia de entrada para construir la salida en cada paso. La arquitectura de atención en los transformers permite, así, una manipulación efectiva de dependencias a largo plazo en secuencias, reduciendo la complejidad computacional respecto a RNNs o CNN. (Reddy, S., 2023).

Un aspecto clave de los transformers es su capacidad para el aprendizaje por transferencia, que consiste en entrenar modelos en grandes cantidades de datos no etiquetados, como Wikipedia, y luego ajustar estos modelos para tareas específicas. Modelos derivados de esta arquitectura, como *Generative pre-trained transformer (GPT)* y *BERT*, han sido pioneros en este enfoque, permitiendo evitar el entrenamiento desde cero para cada tarea (Tunstall et al., 2022).

La capacidad de los transformers para aprender dependencias a larga distancia con una menor complejidad computacional los ha hecho superiores a enfoques anteriores, como las redes recurrentes (RNNs) o convolucionales (ConvS2S). Esto ha sido demostrado en tareas de traducción automática, donde los transformers lograron un rendimiento superior en métricas como BLEU, estableciendo nuevos estándares en la calidad de la traducción con un menor costo computacional (Vaswani et al., 2017).

2.1.2. Modelos basados en transformers

La arquitectura transformer a diferencia de las RNNs, procesa la secuencia completa de palabras de forma simultánea. Esta capacidad permite al modelo comprender mejor el contexto global de cada palabra, lo que facilita una mayor paralelización y reduce significativamente los tiempos de entrenamiento (Plaza Del Arco, 2023). Su principal innovación es el mecanismo de atención, que capta las relaciones contextuales entre palabras o sub-palabras en un texto (Vaswani et al., 2017). A continuación se presentan

algunos de los modelos basados en transformers más importantes.

2.1.2.1. BERT

BERT, propuesto por Devlin et al. en 2019, se entrena en dos fases: preentrenamiento y ajuste fino. El modelo fue preentrenado principalmente en inglés utilizando grandes corpus no etiquetados, como el BooksCorpus y Wikipedia. En el preentrenamiento, el modelo utiliza tareas como el Modelado de Lenguaje Enmascarado (MLM) y la Predicción de la Siguiete Oración (NSP) para aprender de grandes corpus no etiquetados. En el ajuste fino, BERT se adapta a tareas específicas con datos etiquetados, reduciendo así el tiempo y los recursos necesarios en comparación con modelos anteriores (Devlin et al., 2018).

Plaza Del Arco (2023) destaca que BERT ofrece ventajas como:

- Mejora en la comprensión contextual mediante un avanzado mecanismo de atención.
- Superioridad en rendimiento frente a métodos previos.
- Capacidad de paralelización para manejar modelos más grandes y grandes volúmenes de texto.
- Menor necesidad de datos anotados gracias al preentrenamiento.

2.1.2.2. RoBERTuito

RoBERTuito es un modelo transformer a gran escala entrenado con tuits en español, que ha demostrado superar a otros modelos preentrenados en español en diversas tareas de clasificación en X (antes Twitter). Basado en la arquitectura RoBERTa, RoBERTuito utiliza el objetivo de MLM y fue entrenado con un corpus de aproximadamente 500 millones de tuits, mayormente en español (92%), y con una menor representación de otros idiomas. El preprocesamiento de datos, esencial para manejar el ruido en los tuits, incluye la normalización de caracteres, el reemplazo de manejadores de usuarios y hashtags por tokens especiales, y la conversión de emojis a sus representaciones textuales. En comparación con otros modelos como BETO, RoBERTa-BNE y BERTin, RoBERTuito ha demostrado un rendimiento superior

en tareas de análisis de sentimiento, emociones, ironía y discurso de odio. Aunque la evaluación de RoBERTuito se ha centrado en tareas de clasificación debido a la falta de conjuntos de datos para otras tareas, se observaron resultados positivos (Pérez et al., 2021). Como se puede ver en la Tabla 1 se desarrollaron versiones cased, uncased y deaccento del modelo, con la versión uncased mostrando un rendimiento ligeramente mejor. Por lo anterior, se decidió utilizar el modelo " pysentimiento/robertuito-base-uncased" de HuggingFace para nuestro trabajo.

Tabla 1. Resultados de evaluación para las tareas de clasificación en español: detección de discurso de odio, análisis de sentimientos, análisis de emociones y detección de ironía. Los resultados se expresan como la media de la puntuación Macro F1 de 10 ejecuciones de los experimentos de clasificación (Pérez et al. (2021)).

Modelo	Odio	Sentimiento	Emoción	Ironía	Calificación
RoBERTuito_{uncased}	80.1	70.7	55.1	73.6	69.9
RoBERTuito _{deacc}	79.8	70.2	54.3	74.0	69.6
RoBERTuito _{cased}	79.0	70.1	51.9	71.9	68.2
RoBERTa	76.6	66.9	53.3	72.3	67.3
BERT _{in}	76.7	66.5	51.8	71.6	66.7
BETO _{cased}	76.8	66.5	52.1	70.6	66.5
BETO _{uncased}	75.7	64.9	52.1	70.2	65.7

2.2. Explicabilidad

La explicabilidad se refiere a la capacidad de un modelo de inteligencia artificial para proporcionar razones claras y comprensibles sobre cómo llegó a una decisión o predicción (Pombo Nartallo, V., 2024). En el procesamiento del lenguaje natural (PLN), la explicabilidad es esencial para garantizar que los modelos sean transparentes y confiables, permitiendo a los usuarios entender cómo las diferentes partes de un texto influyen en la salida del modelo. Esto es especialmente importante en aplicaciones críticas, como la detección de discurso de odio o la asistencia médica, donde los errores pueden tener graves consecuencias.

Por ejemplo, en el caso de la identificación de discurso de odio, un modelo puede equivocarse al clasificar una expresión sarcástica como ofensiva. Si el modelo es interpretable, es posible examinar qué palabras o expresiones específicas influyeron en esa decisión incorrecta, lo que facilita ajustar el modelo o su interpretación. En aplicaciones médicas, como el apoyo en diagnósticos, es fundamental que el modelo pueda explicar por qué recomendó una determinada condición basándose en los síntomas proporcionados, lo que brinda a los médicos la confianza necesaria para utilizar la herramienta en su proceso de toma de decisiones.

Los mecanismos de atención, en particular, juegan un papel crucial en la explicabilidad al permitir que los modelos identifiquen y ponderen las palabras o frases más relevantes en un texto. A través de la atención, los modelos pueden proporcionar una justificación más clara sobre qué partes del texto fueron más influyentes en la toma de decisiones. A continuación, se describen tres enfoques clave en la explicabilidad.

2.2.1. Valores de Atención

Los valores de atención permiten a los modelos de PLN centrarse en palabras o frases específicas dentro de un texto, asignando una mayor importancia a aquellas que son más relevantes para la tarea en cuestión. Estos valores se derivan de los mecanismos de atención que distribuyen diferentes pesos a las palabras, reflejando su influencia en la predicción del modelo.

Los modelos basados en atención, como las Redes de Atención Jerárquica (HAN), aprovechan esta técnica para construir representaciones más interpretables de textos complejos. En el caso de HAN, se utilizan mecanismos de atención en dos niveles: el nivel de palabra y el nivel de oración. Esto permite que el modelo identifique, dentro de cada oración, qué palabras son las más importantes y, posteriormente, agregue esta información para determinar qué oraciones tienen mayor peso en la clasificación de un documento completo.

2.2.2. Explicabilidad usando redes de atención jerárquicas

La Red de Atención Jerárquica (HAN) está diseñada para capturar la estructura jerárquica de los documentos. Esta arquitectura aborda dos aspectos fundamentales. En primer lugar, dado que los documentos tienen una organización jerárquica (las palabras forman oraciones y las oraciones conforman un documento), HAN construye una representación del documento comenzando por las representaciones de las oraciones y luego integrándolas para formar la representación completa del documento.

En segundo lugar, el modelo reconoce que diferentes palabras y oraciones poseen distintos niveles de importancia que dependen del contexto. Para ello, incorpora mecanismos de atención en dos niveles: a nivel de palabra y a nivel de oración. Estos mecanismos permiten ajustar la atención a palabras y

oraciones individuales durante la construcción de la representación del documento. Por ejemplo, en una revisión de Yelp, se identifican las oraciones y palabras más relevantes para la predicción de la calificación mediante el uso de atención. Esto no solo mejora el rendimiento del modelo, sino que también ofrece información valiosa sobre cómo cada palabra y oración contribuyen a la decisión de clasificación.

A diferencia de enfoques anteriores que utilizan filtros de tokens sin tener en cuenta el contexto, esta arquitectura emplea el contexto para identificar secuencias de tokens relevantes. En la evaluación de su rendimiento en comparación con otras arquitecturas comunes, HAN muestra una mejora significativa en seis conjuntos de datos, destacándose especialmente en la clasificación de documentos (Yang et al., 2016).

2.2.3. ChatGPT

Desarrollado por OpenAI, es un modelo de lenguaje avanzado basado en la arquitectura Generative Pre-training Transformer (GPT), que se destaca por su capacidad para generar lenguaje natural coherente y comprensible (Singh et al., 2023; Diego Olite et al., 2023). Más allá de sus múltiples aplicaciones, ChatGPT ha demostrado ser una herramienta valiosa para mejorar la "explicabilidad" de los sistemas de inteligencia artificial (IA). La explicabilidad en IA se refiere a la capacidad de un sistema para describir de manera clara y comprensible cómo toma decisiones o genera resultados (?). En este contexto, ChatGPT puede actuar como un puente entre modelos de IA complejos y los usuarios finales, traduciendo procesos técnicos en explicaciones accesibles. Esto es especialmente útil en áreas donde la transparencia y la confianza son cruciales, como la salud, la justicia y las finanzas.

Una de las principales aplicaciones de ChatGPT en la explicabilidad es su capacidad para interpretar modelos complejos. Por ejemplo, puede generar explicaciones claras sobre cómo funcionan las redes neuronales profundas o los sistemas de aprendizaje por refuerzo. Esto permite que usuarios no especializados comprendan cómo un sistema de IA procesa datos para generar sus resultados. Además, ChatGPT puede adaptar el nivel de detalle y tecnicismo según la audiencia, ofreciendo explicaciones detalladas para especialistas o resúmenes simples para usuarios finales.

Otro uso relevante de ChatGPT es su capacidad para asistir en el análisis de errores y la depuración de sistemas de IA. Ante resultados inesperados, ChatGPT puede proponer hipótesis sobre posibles causas, facilitando la identificación de problemas en los datos o en la configuración del modelo. Asimismo, en

sistemas de apoyo a la toma de decisiones, ChatGPT puede justificar recomendaciones, explicando qué datos fueron considerados y por qué ciertas opciones fueron priorizadas. Esto contribuye a aumentar la confianza del usuario en las decisiones automatizadas.

Sin embargo, el uso de ChatGPT en la explicabilidad también presenta desafíos. Por un lado, su conocimiento sobre modelos específicos está limitado a generalizaciones aprendidas durante su entrenamiento, lo que implica que no puede analizar directamente los procesos internos de otros sistemas de IA. Por otro lado, puede generar explicaciones plausibles pero incorrectas debido a sesgos en los datos de entrenamiento o a la falta de información actualizada. Además, es importante considerar que sus respuestas pueden reflejar sesgos inherentes a los datos con los que fue entrenado, lo que podría afectar la neutralidad de sus explicaciones (OpenAI, 2022).

En la figura 2 se ilustra el flujo general de ChatGPT, lo que permite entender su potencial como herramienta de explicabilidad en sistemas de IA.

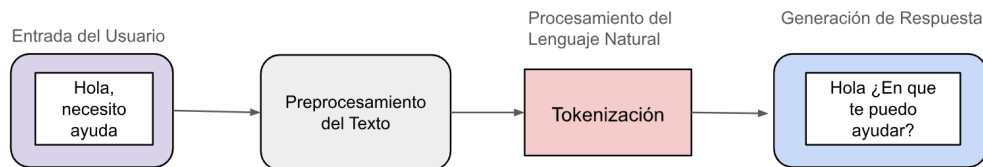


Figura 2. Funcionamiento de ChatGPT.

En conclusión, ChatGPT representa una herramienta innovadora para promover la explicabilidad en sistemas de IA, al traducir procesos técnicos complejos en términos accesibles para usuarios finales y expertos. Si bien tiene limitaciones inherentes a su diseño, su uso estratégico puede aumentar la transparencia, la confianza y la comprensión de las decisiones automatizadas. Por lo tanto, su implementación en contextos donde la explicabilidad es crítica debe realizarse de manera supervisada, aprovechando sus fortalezas y mitigando sus limitaciones (OpenAI, 2022; Singh et al., 2023).

2.3. Resumen del capítulo

Se presentó la arquitectura de transformers, un modelo revolucionario que mejora la eficiencia y precisión en tareas de PLN mediante mecanismos de atención que permiten procesar secuencias en paralelo, capturando dependencias de largo alcance. Modelos basados en transformers, como BERT y RoBERTa, han demostrado ser particularmente efectivos en tareas de clasificación de texto, con aplicaciones en la detec-

ción de discurso de odio, análisis de sentimientos y más. En particular, se destaca el uso de RoBERTuito, entrenado específicamente en tuits en español, como el modelo utilizado en esta investigación.

Finalmente, se exploran aspectos de la explicabilidad en modelos de PLN, como las HAN. Se discute la relevancia de herramientas basadas en inteligencia artificial, como ChatGPT, resaltando tanto sus capacidades como sus limitaciones, tales como sesgos y generación de respuestas incorrectas.

Capítulo 3. Trabajo relacionado

En este capítulo se revisan los trabajos más recientes en la detección de sexismo en redes sociales, con un enfoque en las técnicas de procesamiento del lenguaje natural (PLN). Luego, se analiza en detalle la competencia EXIST¹, la cual se ha consolidado como un foro de evaluación para enfoques innovadores y desafiantes en la detección de contenido sexista en diversos idiomas, con especial atención en el español. Posteriormente, se analizan los trabajos con el mejor desempeño al utilizar el conjunto de datos de la competencia.

3.1. Competencia EXIST: Ediciones 2023 y 2024

Como se mencionó en capítulos previos, el sexismo es un problema persistente en las interacciones en línea, con redes sociales como X (antes Twitter), Facebook e Instagram sirviendo tanto como plataformas para la denuncia de abusos y discriminación, como para la difusión de discursos sexistas. Movimientos como #MeToo y #Time'sUp han visibilizado experiencias de acoso y desigualdad, mientras que, paralelamente, estos mismos espacios digitales han facilitado la proliferación de comportamientos sexistas. Esta dualidad ha generado la necesidad de desarrollar herramientas automáticas capaces de detectar, analizar y medir la presencia de sexismo en las redes sociales.

En este contexto, la competencia EXIST (sEXism Identification in Social neTworks) ha sido una de las iniciativas más influyentes. Organizada dentro de los laboratorios de CLEF² desde su primera edición en 2021, EXIST tiene como objetivo fomentar el desarrollo de modelos automáticos para la identificación del sexismo en redes sociales. Esta competencia abarca desde formas explícitas de misoginia hasta expresiones más sutiles de discriminación de género. La competencia trabaja con un conjunto de datos multilingüe, incluyendo tuits tanto en inglés como en español, lo que amplía su alcance y permite a los modelos desarrollados detectar y analizar el sexismo en distintas lenguas y contextos culturales. Esta diversidad lingüística plantea un reto adicional para los participantes, ya que deben crear sistemas capaces de abordar matices culturales y lingüísticos específicos en la identificación de actitudes sexistas.

Aunque existen otros estudios relacionados con el análisis de sexismo en redes sociales, estos suelen diferir en aspectos clave, como los conjuntos de datos empleados, las métricas utilizadas o los enfoques metodológicos, lo que limita su aplicabilidad. En contraste, los trabajos relacionados con EXIST abordan

¹<https://nlp.uned.es/exist2024/>

²<https://clef2024.imag.fr/>

de manera directa los mismos desafíos y requisitos específicos que enfrenta este proyecto, proporcionando un marco sólido y relevante para sustentar nuestras contribuciones.

En las siguientes subsecciones, se describirán con mayor detalle las tareas propuestas en la competencia EXIST 2023 y 2024, incluyendo su definición, ejemplos representativos y los objetivos específicos de cada una. Además, se explicarán las métricas de evaluación utilizadas para medir el desempeño de los sistemas participantes, con énfasis en el enfoque Hard-Hard y el uso de la métrica F1 score macro. Esto permitirá comprender los desafíos que enfrentan los modelos en la detección y clasificación del sexismo en redes sociales, así como los criterios que determinan su eficacia en un entorno multilingüe y culturalmente diverso.

3.1.1. Identificación de sexismo

Consiste en una tarea de clasificación binaria, en la cual los sistemas deben determinar si un tuit dado contiene expresiones o comportamientos sexistas. Esto incluye mensajes que son sexistas en sí mismos, que describen situaciones sexistas o que critican comportamientos sexistas (Plaza et al., 2023). Ejemplos de ellos son:

- **Sexista:** Mujer al volante, tenga cuidado!
- **No Sexista:** Alguien me explica que zorra hace la gente en el cajero que se demora tanto.

El primer ejemplo, “Mujer al volante, tenga cuidado!”, se considera sexista porque perpetúa un estereotipo de género que asocia la habilidad de conducir con el género, sugiriendo que las mujeres son inherentemente menos competentes al volante. Este tipo de afirmación refuerza prejuicios negativos hacia las mujeres y contribuye a la discriminación de género. En contraste, el segundo ejemplo, “Alguien me explica que zorra hace la gente en el cajero que se demora tanto”, aunque emplea una palabra potencialmente ofensiva, no es sexista porque no refuerza ni perpetúa estereotipos de género ni discrimina a un género en particular. La expresión puede ser grosera, pero su intención no está dirigida a discriminar o criticar con base en el género, sino que es una queja general sobre una situación específica.

3.1.2. Intención de la fuente

Una vez clasificado un mensaje como sexista, la segunda tarea busca categorizarlo según la intención del autor, proporcionando así una visión del papel que juegan las redes sociales en la emisión y difusión de mensajes sexistas. Esto implica categorizar la intención del autor, ya sea que la intención sea directa, de reporte o de crítica (Plaza et al., 2023). A continuación, se incluyen algunos ejemplos representativos de cada una de estas categorías para ilustrar sus características principales:

- **Directa:** La intención era escribir un mensaje que, por sí solo, fuera sexista o incitara al sexismo, como por ejemplo: “Las mujeres no deberían programar. . . quizá ser influencers o creadoras. . . es su fortaleza natural.”
- **Reportada:** La intención es informar y compartir una situación sexista vivida por una mujer o por varias mujeres, ya sea en primera o tercera persona, como por ejemplo: “Me adormezco en el metro, abro los ojos sintiendo algo raro: la mano del hombre sentado a mi lado en mi pierna #SquealOnYourPig.”
- **Juicio:** La intención era juzgar, ya que el tuit describe situaciones o comportamientos sexistas con el propósito de condenarlos, como por ejemplo: “Como de costumbre, la mujer fue quien dejó su trabajo por el bienestar de la familia...”

3.1.3. Categorización de tipo de sexismo

La categorización del sexismo en los tuits se enfoca en clasificar los tuits sexistas en categorías específicas como ideología e inequidad, estereotipos y dominancia, cosificación, violencia sexual, misoginia y violencia no sexual (Plaza et al., 2023). A continuación, se presentan ejemplos representativos de cada categoría para ilustrar cómo se manifiestan estas formas de sexismo en los mensajes analizados:

- **Ideología y desigualdad:** El texto desacredita el movimiento feminista, rechaza la desigualdad entre hombres y mujeres o presenta a los hombres como víctimas de la opresión de género, como por ejemplo: “Mi hermana y mi madre se burlan de mí por defender los derechos de todos y me acaban de llamar feminazi.”

- **Estereotipos y dominancia:** El texto perpetúa estereotipos de género, sugiriendo que las mujeres son más aptas para roles tradicionales (madre, esposa, cuidadora) y menos para ciertas tareas, o que los hombres son superiores a ellas, como por ejemplo: “@Paula2R @faber_acuria A las mujeres hay que amarlas... solo eso... Nunca las entenderás.”
- **Cosificación:** El texto reduce a las mujeres a objetos, negando su dignidad y promoviendo estándares de belleza y roles de género tradicionales que las hipersexualizan, como por ejemplo: “Pareces una puta con ese pantalón” - Mi hermano de 13 cuando me vio con un pantalón de cuero.
- **Violencia Sexual:** Se realizan propuestas sexuales, demandas de favores sexuales o se da lugar a acoso de índole sexual, incluyendo violación o agresión sexual; como por ejemplo: “#MeToo Estas 4 no han conseguido su objetivo. El juez estima que se abrieron de patas <https://t.co/GSHiiwqY6A> ¡Cuánta lagartona hay en este #metoo!”
- **Misoginia y Violencia No Sexual:** El texto transmite hostilidad y agresión hacia las mujeres, como por ejemplo: “Las mujeres de hoy en día te enseñan a querer... estar soltero.”

3.1.4. Métricas de evaluación

Para la evaluación de los sistemas de clasificación participantes en las tres tareas de EXIST, se utilizaron dos enfoques distintos, según el tipo de resultados generados por los sistemas y las etiquetas del *ground truth* (referencia de verdad). Estos enfoques son:

- **Evaluación Hard-Hard:** Aplica cuando tanto el sistema como el *ground truth* generan salidas estrictas o “duras” (clasificaciones únicas y definidas).
- **Evaluación Soft-Soft:** Se utiliza cuando tanto el sistema como el *ground truth* producen salidas probabilísticas o “suaves”, donde se asignan probabilidades a distintas categorías.

Dado que en este trabajo consideramos exclusivamente la evaluación Hard-Hard, damos mayor detalle del proceso de obtención de las categorías “duras” y de la métrica de evaluación utilizada en ese caso. Las etiquetas del *ground truth* se obtuvieron a partir de los votos de los anotadores humanos, aplicando un umbral probabilístico específico para cada tarea:

- En la tarea 1, se selecciona la clase anotada por más de 3 personas. Por ejemplo, si las etiquetas obtenidas son [“SI”, “NO”, “NO”, “NO”, “SI”, “NO”], la etiqueta “dura” en este caso sería “NO”.
- En la tarea 2, se elige la clase anotada por más de 2 personas o aquella con mayor consenso. Por ejemplo, si tienen las etiquetas [“JUICIO”, “DIRECTO”, “REPORTE”, “REPORTE”, “REPORTE”, “JUICIO”], su etiqueta “dura” sería “REPORTE”.
- En la tarea 3 (multi-etiqueta), se seleccionan las clases anotadas por más de 1 persona. Por ejemplo, si se tienen las etiquetas [[“COSIFICACIÓN”, [“COSIFICACIÓN”, “VIOLENCIA-SEXUAL”], [“COSIFICACIÓN”, [“VIOLENCIA-SEXUAL”], [“COSIFICACIÓN”, [“MISOGINIA-VIOLENCIA-NO-SEXUAL”]]], la etiqueta “dura” en este caso sería [“COSIFICACIÓN”, “VIOLENCIA-SEXUAL”].

Nosotros nos enfocamos exclusivamente en la evaluación Hard-Hard, usando la métrica F1 score macro para evaluar salidas categóricas duras y compararlas con el *ground truth* obtenido de los anotadores humanos. Los detalles específicos sobre el cálculo de la métrica F1 score macro se presentan en el Anexo A.

3.2. Revisión de la literatura

El análisis de las competencias *EXIST* 2023 y 2024 pone en evidencia los avances significativos en las técnicas de PLN para la identificación, clasificación y categorización del sexismo en redes sociales. Estas competencias, al ofrecer datos multilingües y tareas complejas, han incentivado la innovación en enfoques y herramientas para abordar problemas lingüísticos y sociales en diferentes contextos. En este capítulo, se organiza la información de forma estructurada en torno a las principales técnicas empleadas por los equipos destacados, obtenidos para el idioma español y desde la perspectiva Hard-Hard. Destacando patrones comunes, innovaciones clave y los resultados obtenidos en cada tarea.

3.2.1. Principales técnicas y estrategias empleadas

3.2.1.1. Preprocesamiento y preparación de datos

El preprocesamiento de datos fue una etapa crucial en los enfoques de los equipos participantes. Las técnicas aplicadas permitieron estandarizar los tuits, aumentar la calidad de los datos y mejorar el

rendimiento de los modelos.

En 2024, el equipo *NYCU-NLP* implementó un preprocesamiento exhaustivo que incluyó la eliminación de elementos irrelevantes como nombres de usuario, URLs, emojis y hashtags. También llevaron a cabo una estandarización textual, convirtiendo todos los tuits a minúsculas para asegurar uniformidad en el análisis. Además, emplearon técnicas de enriquecimiento como *back-translation*, utilizando la API de Google Translate para traducir tuits del inglés al español y de regreso al inglés, generando variaciones sutiles que incrementaron la robustez del modelo. Otro método destacado fue el *AEDA*, una técnica que segmenta oraciones y agrega puntuaciones aleatorias para diversificar los datos sin alterar su significado (Fang et al., 2024).

Por otro lado, el equipo *roh-neil* (2023) adoptó un enfoque más simple pero efectivo, reemplazando menciones de usuarios con "@USERz" y URLs con "#HTTPURL", manteniendo la estructura y matices originales de los tuits. Este enfoque permitió conservar las particularidades del lenguaje social sin perder información relevante. Adicionalmente, emplearon datos de la edición *EXIST* 2021 traducidos al español y al inglés utilizando la herramienta *googletrans*, lo cual resultó fundamental para expandir su conjunto de entrenamiento y mejorar la generalización del modelo (Koonireddy & Adel, 2023).

3.2.1.2. Modelos basados en transformers

Los modelos basados en transformers representaron el núcleo de los enfoques de la mayoría de los equipos, destacando por su capacidad de manejar datos multilingües y su flexibilidad para adaptarse a tareas específicas.

El equipo *NYCU-NLP* (2024) utilizó modelos como *DeBERTa-v3* y *XLM-RoBERTa*, combinados con la técnica de *Hard Parameter Sharing*, que permite compartir capas ocultas entre tareas relacionadas, optimizando el uso de recursos computacionales y reduciendo el sobreajuste. Esta técnica facilitó que los modelos generalizaran mejor entre tareas como la identificación del sexismo y la clasificación de intenciones (Fang et al., 2024).

Por su parte, el equipo *ABCD* empleó modelos grandes como *Llama 2* y *T5*, que fueron ajustados utilizando *prompt engineering*, junto con modelos más pequeños como *XLM-RoBERTa*, los cuales fueron afinados directamente con los datos procesados. Esta arquitectura en componentes permitió reflejar mejor

las diferencias en las percepciones de los anotadores y manejar eficientemente la información de cada grupo. Además, utilizaron técnicas avanzadas como *LoRA (Low-Rank Adaptation)*, que les permitió ajustar *Llama 2* con menor costo computacional y alta precisión (Quan & Thin, 2024).

El equipo *roh-neil (2023)* optó por modelos como *XLM-RoBERTa-Large*, que demostraron ser superiores a modelos específicos de idioma como *RobertaTuito* o *BERTweet-large* en tareas multilingües. Este equipo realizó ajustes finos utilizando *Optuna* para optimizar hiperparámetros, lo que mejoró significativamente la precisión del modelo en la clasificación de tuits sexistas (Koonireddy & Adel, 2023).

3.2.1.3. Uso de Modelos de Lenguaje Grandes para explicabilidad

Los Modelos de Lenguaje Grandes (LLMs) se utilizaron tanto para clasificación como para generación de explicaciones, ampliando su papel más allá de las tareas tradicionales de PLN.

El equipo *CIMAT-CS-NLP (2023)* utilizó *Gemini-1.0-pro* para implementar una clasificación *zero-shot* a través de prompts refinados, diseñados para mejorar la comprensión y precisión del modelo. Los prompts evolucionaron desde versiones simples que pedían respuestas binarias (“Sí” o “No”) hasta configuraciones más complejas que simulaban a expertos en sexismo. Este enfoque permitió obtener resultados competitivos sin necesidad de un entrenamiento extenso (Tavarez-Rodríguez et al., 2024).

Además, el equipo *CIMAT-GTO* empleó un modelo como *Llama 3* para generar razonamientos explicativos, tanto positivos como negativos, sobre si un tuit era sexista. Estos razonamientos no solo servían para la clasificación, sino que también enriquecían los datos al generar análisis detallados que luego se combinaban con un modelo *XLM-RoBERTa* preentrenado (Villarreal-Haro et al., 2024).

3.2.1.4. Incorporación de información de anotadores

Un elemento innovador en *EXIST 2024* fue la inclusión de metadatos sobre los anotadores, lo cual mejoró la precisión de los modelos al capturar sesgos o patrones específicos en las anotaciones.

El equipo *NYCU-NLP* utilizó atributos demográficos de los anotadores (género, edad, etnicidad) codificados mediante *one-hot encoding* y procesados con capas de embeddings para generar representaciones

densas que se integraron al modelo. Esto permitió ajustar las predicciones según las características de los anotadores, mejorando la interpretación de los datos (Fang et al., 2024).

Por su parte, el equipo *ABCD* implementó modelos independientes para cada grupo de anotadores. Estos modelos fueron entrenados individualmente y luego sus predicciones se fusionaron utilizando umbrales probabilísticos, reflejando de forma más precisa las decisiones colectivas (Quan & Thin, 2024).

3.3. Comparación de desempeño y enfoques en las competencias EXIST 2023 y 2024

En la *Tabla 2* se presenta una comparación detallada de los resultados obtenidos en las competencias *EXIST* 2023 y 2024, destacando los equipos con mejor desempeño en cada tarea, sus posiciones en el ranking, las corridas específicas y las métricas *F1* alcanzadas. Esta comparación permite observar las mejoras progresivas en las técnicas de clasificación y ajuste fino de modelos en la detección de discursos sexistas, así como las variaciones en rendimiento entre ambos años.

Se destacan los equipos con mejor desempeño en cada tarea, especificando sus posiciones en el ranking, las configuraciones implementadas y las métricas de *F1* alcanzadas. Este análisis permite evaluar el progreso técnico en la clasificación y ajuste fino de modelos de lenguaje para la detección de discursos sexistas, además de revelar las variaciones de rendimiento entre ambos años.

Tabla 2. Tabla Comparativa EXIST 2023 y 2024

Tarea	Ranking	2023			2024		
		Equipo	Corrida	F1	Equipo	Corrida	F1
Tarea1	1	Mario	1	0.8384	NYCU_NLP	1	0.8238
	2	Mario	3	0.8387	ABCD Team	1	0.8097
	3	Mario	2	0.83	CIMAT-CS-NLP	2	0.8148
	4	roh_neil	1	0.8054	CIMAT-GTO	3	0.8123
	5	roh_neil	2	0.8054	NYCU-NLP	3	0.8075
Tarea2	1	Mario	2	0.6059	NYCU-NLP	1	0.5757
	2	roh_neil	1	0.5855	ABCD Team	1	0.5808
	3	roh_neil	2	0.5855	NYCU-NLP	3	0.5697
	4	UniBo	2	0.5654	NYCU-NLP	2	0.5723
	5	AIT_FHSTP	1	0.5414	CUET-SSTM	1	0.5779
Tarea3	1	roh_neil	1	0.6431	ABCD Team	1	0.6219
	2	roh_neil	2	0.6431	ABCD Team	3	0.6247
	3	AIT_FHSTP	1	0.5995	NYCU-NLP	3	0.6228
	4	Mario	3	0.5578	NYCU-NLP	1	0.6206
	5	UniBo	2	0.6055	NYCU-NLP	2	0.5968

Este capítulo revisa los enfoques más destacados en la detección de sexismo en redes sociales, con un énfasis especial en las ediciones 2023 y 2024 de EXIST. Se examinan las estrategias implementadas por los equipos mejor clasificados, quienes utilizaron modelos de lenguaje avanzados y técnicas de ajuste fino para abordar la complejidad del discurso sexista en español. Los resultados presentados en la *Tabla 3* demuestran cómo el uso de arquitecturas avanzadas como GPT-NeoX, XLM-RoBERTa y DeBERTa-v3, combinado con técnicas de ajuste fino específicas para tareas y dominios, ha mejorado significativamente la precisión en la identificación y categorización de discursos sexistas.

Por ejemplo, en la edición 2024, el equipo NYCU-NLP alcanzó un *F1* de 0.8238 en la Tarea 1 mediante un enfoque que integró preprocesamiento extensivo y estrategias de optimización específicas. Si bien los avances son notables, los resultados también evidencian la necesidad de desarrollar enfoques más robustos, particularmente en la integración de razonamientos explicativos y en la transparencia de los modelos. Por ejemplo, en la edición 2024, los enfoques basados en razonamientos (como los explorados por CIMAT-GTO con Llama 3) mostraron mejoras, pero aún enfrentan limitaciones en términos de generalización y precisión.

Los equipos participantes aplicaron una amplia gama de métodos y modelos, desde arquitecturas generales como LLaMA y GPT-NeoX, hasta modelos optimizados como XLM-RoBERTa y mDeBERTa. Además, exploraron estrategias innovadoras, entre ellas:

- La combinación de etiquetas blandas y duras para mejorar la robustez del aprendizaje.
- Preentrenamientos específicos para datos de Twitter, adaptando los modelos al contexto de las redes sociales.
- Uso de embeddings enriquecidos con información de sentimiento y toxicidad para capturar matices adicionales del discurso.
- La integración de razonamientos explicativos en el análisis del discurso, como en los enfoques de CIMAT-GTO, que utilizaron Llama 3 para generar razonamientos positivos, negativos y comparativos. Estos razonamientos permiten una mejor comprensión e interpretación del contexto del discurso sexista.
- Enfoques orientados a mejorar la transparencia de los modelos, como el uso de métodos de optimización y fine-tuning específicos que facilitan el análisis interpretativo de los resultados.

Tabla 3. Métodos y resultados de los equipos participantes en las tareas del EXIST 2023 (Plaza et al., 2023) y 2024 (Plaza et al., 2024).

Año	Equipo	Participación	Modelos utilizados	Métodos y enfoques principales
2023	Team Mario	Tareas 1, 2 y 3	GPT-NeoX (inglés), BERTIN-GPT-J-6B (español)	Organización en cascadas, fine-tuning con datos in-domain, boosting con datasets de hate speech (Hugging Face), fine-tuning específico por tarea.
2023	roh-neil	Tareas 1, 2 y 3	xlm-roberta-large-twitter	Optuna para optimización de parámetros, aumento de datos con datasets de ediciones previas de EXIST.
2023	UniBo	Tareas 1, 2 y 3	RobertaHate, EmoBerta	Traducción con Google API (español a inglés), incorporación de emociones como características adicionales a embeddings de RobertaHate.
2023	AIT_FHSTP	Tareas 1, 2 y 3	XLM-RoBERTa, Random Forest	Embeddings de análisis de sentimientos y toxicidad, PCA, entrenamiento de Random Forest, predicción multi-etiqueta en Tarea 3.
2024	NYCU-NLP	Tareas 1, 2 y 3	DeBERTa-v3, xlm-RoBERTa	Preprocesamiento extensivo (remoción de irrelevantes, back-translation, AEDA), integración de información del anotador (género, edad, etnia), Hard Parameter Sharing, Round to Closest Value.
2024	ABCD	Tareas 1, 2 y 3	Llama 2, T5, XLM-RoBERTa	División por grupos de anotadores, prompt engineering, fine-tuning por subgrupos, estructura jerárquica para Tareas 2 y 3 (solo sexistas en Tarea 1).
2024	CIMAT-CS-NLP	Tarea 1 y 2	Gemini API (zero-shot), XLM-RoBERTa, mBERT, Twitter-XLM-RoBERTa	Clasificación zero-shot con prompting, fine-tuning supervisado de modelos transformers, integración de resultados por votación, input generado y mejores modelos.
2024	CIMAT-GTO	Tarea 1	Llama 3, XLM-RoBERTa	Generación de razonamientos con Llama 3, integración de razonamientos al texto, clasificación con XLM-RoBERTa pre-entrenado. Exploración de razonamientos positivos, negativos y comparativos.

En conjunto, estos enfoques han permitido avanzar significativamente en la capacidad de los modelos para entender, detectar y categorizar el discurso sexista, como se evidencia en el análisis de las competencias EXIST. El uso de técnicas avanzadas de PLN, incluyendo modelos transformadores, LLMs y estrategias híbridas, ha impulsado estos progresos, mientras que la incorporación de razonamientos explicativos añade una dimensión interpretativa crucial al proceso. Estas innovaciones no solo mejoran la precisión de los modelos, sino que también abren nuevas oportunidades para una moderación de contenido más efectiva y un análisis social más profundo en este contexto crítico. Sin embargo, persisten desafíos, particularmente en el diseño de sistemas que, además de ser más precisos, sean transparentes y explicativos, lo que resulta esencial para abordar esta problemática de manera integral.

Capítulo 4. Análisis de Percepciones sobre el Sexismo

En este capítulo, analizamos el proceso de etiquetado del contenido sexista en redes sociales, particularmente en X (antes Twitter), destacando las diferencias en las percepciones correspondientes al género y edad de los anotadores. Este enfoque busca explorar cómo estas diferencias influyen en la identificación y clasificación inicial del contenido.

Como se presentó en el capítulo previo, el dataset de EXIST(Plaza et al., 2023), creado con el propósito de investigar el lenguaje ofensivo y sexista en las redes sociales, aborda un desafío central en el campo del PLN: el sesgo en los datos. Este problema ha cobrado creciente relevancia en los últimos años, especialmente debido al uso masivo de modelos de lenguaje a gran escala que, en muchos casos, no solo reproducen, sino que amplifican los sesgos inherentes en los datos con los que son entrenados y en los algoritmos que procesan dicha información (Roselli et al., 2019).

Entre los tipos de sesgo más críticos se encuentran los relacionados con la edad y el género, ya que estos factores influyen significativamente en cómo se percibe, etiqueta y clasifica el contenido. Por ejemplo, ciertos términos o expresiones pueden considerarse ofensivos o sexistas dependiendo de la perspectiva de quien los evalúe, lo que a menudo refleja diferencias intergeneracionales o basadas en construcciones sociales de género. Estas discrepancias no solo afectan la consistencia en el etiquetado, sino que también repercuten en los modelos entrenados con estos datos, que pueden perpetuar y reforzar estas inequidades en aplicaciones prácticas. Abordar estos sesgos es fundamental para desarrollar sistemas más justos y representativos en el análisis automatizado del lenguaje.

4.1. Conjunto de datos EXIST

El dataset EXIST se centra en mitigar el sesgo introducido durante la selección de datos y el proceso de etiquetado. Para ello, sigue el paradigma de “aprendizaje con desacuerdo” (*Learning with Disagreement*, LwD), lo que permite que los sistemas aprendan de información que puede ser conflictiva o subjetiva, dado que las percepciones sobre el sexismo y otros tipos de lenguaje ofensivo pueden variar según el género, edad o contexto cultural de los anotadores. Este enfoque es particularmente importante en el análisis de expresiones sexistas, ya que lo que una persona percibe como sexista puede no serlo para otra, dependiendo de su trasfondo cultural y social.

El proceso de etiquetado en EXIST fue diseñado específicamente para minimizar el sesgo introducido por

las diferencias sociodemográficas de los anotadores. Las anotaciones fueron realizadas por trabajadores de la plataforma Prolific, seleccionados en función de características demográficas como género (masculino/femenino) y edad (18–22, 23–45 y más de 46 años), con el objetivo de asegurar una diversidad de perspectivas. Cada tuit fue revisado por seis anotadores, lo que permitió recoger una variedad de interpretaciones sobre si el contenido era sexista o no.

Siguiendo este enfoque, EXIST preserva todas las etiquetas proporcionadas por los distintos anotadores, en lugar de una única etiqueta agregada, lo que permite captar la diversidad de interpretaciones y mitigar los sesgos inherentes a las decisiones individuales de los anotadores. Este método sigue las recomendaciones de estudios previos (Akhtar et al., 2021) que destacan la importancia de trabajar con datos etiquetados con desacuerdo para desarrollar sistemas de PLN más equitativos y menos propensos a sesgos.

Entre 2021 y 2022, EXIST ofrecía un conjunto de datos con etiquetas únicas y definitivas para cada tuit. Sin embargo, desde 2023 y continuando en 2024, la tarea evolucionó para asignar seis etiquetas distintas a cada tuit, basadas en las perspectivas de seis perfiles de anotadores. En total, participaron 725 anotadores: 390 para los tuits en español y 335 para los tuits en inglés. Estos perfiles comprenden tres mujeres y tres hombres distribuidos en tres rangos de edad: 18–22, 23–45 y más de 46 años. La segmentación demográfica se diseñó considerando teorías sobre las etapas de la vida y los niveles de interacción digital característicos de cada grupo:

- **Adolescencia y adultez temprana (18–22):** Este grupo se caracteriza por su alta exposición a las redes sociales y entornos digitales, ya que son nativos digitales que han crecido con las redes sociales como una parte central de sus vidas.
- **Adultez temprana y media (23–45):** Este grupo incluye a personas que, aunque son competentes en el uso de herramientas y redes digitales, pueden no tener la misma exposición intensa a las redes sociales que los usuarios más jóvenes. Son usuarios activos, pero tienden a equilibrar de manera distinta su vida digital y offline.
- **Adultez tardía (46+):** Este grupo generalmente tiene un menor nivel de interacción con las plataformas digitales, especialmente con las redes sociales, en comparación con los grupos más jóvenes. Sus perspectivas suelen reflejar una interacción más ocasional o distanciada con el contenido digital, lo cual puede influir en su proceso de etiquetado.

Además, la disponibilidad de usuarios en las plataformas digitales también fue un factor importante para

definir esta segmentación específica. Por último, estos parámetros son fundamentales para garantizar que las diferencias culturales, sociales y personales queden reflejadas en las etiquetas, contribuyendo a un análisis más inclusivo.

Para controlar la calidad de las anotaciones, se aplicaron diversos mecanismos, como el análisis de *outliers*, la evaluación del tiempo dedicado a las tareas y la inclusión de datos de referencia (*ground truth*). Asimismo, mantuvo una comunicación constante con los anotadores para resolver dudas y corregir errores. Las directrices para la anotación, desarrolladas por dos expertos en estudios de género, incluyeron ejemplos detallados de las diferentes categorías y etiquetas, lo que aseguró una comprensión clara de la tarea (Plaza et al., 2024).

4.1.1. Muestreo de datos

El proceso de recopilación de datos para EXIST incluyó una selección balanceada de expresiones sexistas y no sexistas en español e inglés, recopiladas de diversas fuentes como trabajos previos, cuentas de X (antes Twitter), y proyectos como *Everyday Sexism* y diccionarios feministas. Estas fuentes proporcionaron un amplio conjunto de semillas que luego fueron utilizadas para recolectar datos en Twitter. El conjunto de datos resultante se organizó en tres subconjuntos: entrenamiento, desarrollo y prueba. Estos contienen más de 3,200 tuits por idioma para el conjunto de entrenamiento, y alrededor de 500 y 1,000 tuits por idioma en los conjuntos de desarrollo y prueba, respectivamente. La Tabla 4 resume la distribución de etiquetas para las tareas propuestas: clasificación binaria entre expresiones sexistas y no sexistas (Tarea 1), categorización según el tipo de sexismo (Tarea 2) y asignación de etiquetas específicas relacionadas con fenómenos como cosificación o ideología sexista (Tarea 3). Este desglose permite una visión detallada del balance y la diversidad del conjunto de datos.

Para garantizar un equilibrio entre expresiones sexistas y no sexistas, se incluyeron hashtags y frases que, aunque no suelen usarse en contextos sexistas, permitieron obtener una muestra más diversa. El conjunto final de semillas abarcó 183 expresiones en español y 163 en inglés. Con el fin de evitar sesgos relacionados con los autores, se incluyó solo un tuit por cada autor. Asimismo, se eliminaron los tuits que contenían menos de cinco palabras, asegurando una mayor calidad en el conjunto de datos final.

Una característica distintiva de este corpus es la preservación de todas las anotaciones realizadas por seis anotadores distintos (tres mujeres y tres hombres), distribuidos en tres grupos etarios (18–22, 23–45, y

más de 46 años). Cada tuit recibió seis etiquetas independientes, lo que permite capturar la diversidad de interpretaciones. En total, el conjunto de datos incluye más de 21,000 anotaciones individuales, distribuidas equitativamente entre los dos idiomas. Estas etiquetas reflejan las percepciones diversas de los anotadores sobre los mismos textos y sientan las bases para un análisis más profundo de cómo las características demográficas influyen en la evaluación del sexismo.

La Tabla 4 muestra estadísticas detalladas de las etiquetas asociadas al corpus para cada tarea y contexto de evaluación. Esta información permite identificar la amplitud y profundidad del conjunto de datos EXIST 2023.

Tabla 4. Estadísticas de etiquetas duras del conjunto de datos EXIST 2023 (Plaza et al., 2023).

	Entrenamiento			Desarrollo			Prueba		
	ES	IN	Total	ES	IN	Total	ES	IN	Total
Tarea 1									
No Sexista	1634	1733	3367	229	250	479	491	489	980
Sexista	1560	1137	2697	261	194	455	478	349	827
Tarea 2									
Directo	749	545	1294	117	87	204	244	152	396
Reporte	265	194	459	40	35	75	78	45	123
Juicio	228	148	376	55	28	83	78	75	153
Tarea 3									
Cosificación	611	492	1103	88	95	183	194	150	344
Violencia Sexual	359	316	675	74	49	123	99	97	196
Estereotipos y dominancia	810	613	1423	136	105	241	257	187	444
Ideología y desigualdad	632	481	1113	117	95	212	223	161	384
Misoginia y Violencia No Sexual	552	304	856	90	68	158	166	109	275

4.2. Análisis de las diferencias en la percepción del sexismo

La metodología seguida consistió en realizar una comparación de percepciones de sexismo en tuits, analizando de manera independiente los datos de género (hombres y mujeres) y de edad (18-22, 23-45 y 46+ años). El objetivo fue identificar intersecciones o diferencias en la clasificación de tuits como sexistas o no sexistas, evaluando cómo estas percepciones pueden variar tanto por género como por grupo de edad.

Primero, se segmentaron los datos en dos conjuntos separados según el género de los evaluadores (hombres y mujeres). Se buscó identificar si existían coincidencias en la percepción de sexismo en los tuits dentro de cada género. Luego, se procedió de manera similar pero considerando solo los grupos de edad,

dividiendo a los evaluadores en tres rangos: 18-22 años, 23-45 años y 46+ años. Este análisis separado permitió observar cómo las percepciones de sexismo podrían variar exclusivamente por la edad, sin que el género interfiriera en esta parte de la comparación.

Para asegurar una comparación consistente en ambos casos (género y edad), se limitaron los análisis a aquellos tuits que aparecían en todos los subconjuntos definidos, es decir, aquellos tuits compartidos entre todos los grupos de género o edad respectivamente. Esto permitió evaluar únicamente aquellos tuits que tenían representatividad en todos los grupos, evitando así sesgos derivados de contenidos no presentes en todos los segmentos.

Finalmente, se aplicó un análisis comparativo entre los grupos en cada dimensión. Para cada par de grupos de género o edad, se revisaron las etiquetas de clasificación de los tuits para identificar coincidencias y discrepancias en las percepciones. Las coincidencias reflejaron una intersección en las percepciones de sexismo, mientras que las discrepancias destacaron posibles diferencias interpretativas. Esto permitió obtener una visión clara de cómo las interpretaciones de sexismo en los tuits coinciden o difieren, tanto entre hombres y mujeres como entre los distintos grupos de edad.

Esta metodología, basada en análisis separado de género y edad, permitió identificar patrones de interpretación y segmentar las percepciones de sexismo de manera más precisa, ofreciendo una perspectiva enriquecida sobre cómo varía la percepción de contenido sexista en redes sociales según estas dos dimensiones.

Se realizó una evaluación cualitativa del acuerdo y desacuerdo en la percepción del sexismo entre hombres y mujeres, utilizando los tuits etiquetados del corpus EXIST. Aunque este corpus cuenta con anotaciones realizadas por seis personas diferentes, hasta ahora no se había explorado en profundidad cómo varía la percepción del sexismo según las características de los anotadores y anotadoras, como el género y la edad. Este análisis constituye una de las principales contribuciones de nuestra investigación.

4.2.1. Percepción del sexismo por género

De un total de 3660 textos, el 36 % mostró acuerdo entre hombres y mujeres en que los textos son sexistas, mientras que el 39 % coincidió en que no lo eran, como se muestra en la Tabla 5. Sin embargo, en el 12 % de los textos, los hombres consideraron que eran sexistas mientras que las mujeres no, y en el 13 % restante, las mujeres los clasificaron como sexistas mientras que los hombres no lo hicieron. Estas

discrepancias revelan diferencias significativas en la percepción y sensibilidad hacia el contenido sexista entre géneros, acumulando un 25 % de desacuerdo entre ambos grupos.

Tabla 5. Matriz de Confusión de coincidencias y desacuerdos en las etiquetas de sexismo entre Hombres y Mujeres.

		Mujeres	
		Sexismo	No Sexismo
Hombres	Sexismo	1329	434
	No Sexismo	458	1439

Los casos más destacados de desacuerdo ocurren cuando las mujeres etiquetan las publicaciones como “sexistas”, mientras que los hombres las consideran “no sexistas”. Muchos de estos casos giran en torno a opiniones de los hombres sobre las mujeres, presentadas como “hechos”, como se ejemplifica en la Tabla 6.

Tabla 6. Coincidencias y diferencias en el etiquetado según género

Etiqueta	Tuits
femenino: SÍ masculino: SÍ	<ul style="list-style-type: none"> - <i>A partir de los seis años las chicas tienden a pensar que son trabajadoras, pero menos brillantes que sus compañeros, una opinión en la que influyen los estereotipos de #género #coeducacion https://t.co/3obeH4BNMK</i> - <i>Acosar mujeres deliciosas por las diferentes plataformas digitales, además de escribirles cosas sucias, is my passion</i> - <i>la solución de blizzard a los casos de acoso sexualmmujeres taparse por favor</i> - <i>chacha no llores por un tío, y menos si sabes que te puedes tirar a todo su grupo si quieres</i>
femenino: NO masculino: NO	<ul style="list-style-type: none"> - <i>Lorebrou16 arrobaDonko Mierda, me siento vieja cuando decís Gamergate, que bardoXD</i> - <i>Gueisha3 HRodriguezCuan Los políticos acostumbran a hablarle al pueblo como si fueran una manada de estúpidos pero lamanada no hacemos nada por contra-decirlos.</i> - <i>Hoyuelicos Joder que puto asco los tíos así, que luego vendrán gilis a decir notallmen, pero mira, gañán, son inmensa mayoría los que hacen estas mierdas. Es que lo puto reventaba a ese gilipollas joder.</i> - <i>Ya va, pero Eladio abusó con esas canciones de ese álbum</i>
femenino: SÍ masculino: NO	<ul style="list-style-type: none"> - <i>Nadie te va a tratar tan bien como un hombre que te quiere cog*r por primera vez. #BuenosDías</i> - <i>Si Ironman tiene una luz en el pecho... Por qué la mujeres no pueden tener 2???? Se imaginan las mujeres convertidas en punto de referencia??? En un concierto??? De noche en bici??? https://t.co/aCTgOAZPDA</i> - <i>@EstefMolina_ Póngase un cartel,,soy feminazi,,y se ahorra tan brutal conflicto,,q fuerte por farol</i> - <i>@gishel_paola @PaveloRockstar @giov_ Las mujeres no deben opinar, no porque varias lo hagan, tu también andes haciéndolo</i>
femenino: NO masculino: SÍ	<ul style="list-style-type: none"> - <i>No todos los hombres son iguales, siempre llega uno más cul*ro que el anterior.</i> - <i>Quiero un chacalon que tenga coche que me quiera hacer suya que me deje olerle los pies y brincar y que me trate como una señorita una princesa quiero un chacalon de ave tlahuacf calle 11 5588379799</i> - <i>@CalazaRicardo @PatG_ Sr, no todas las mujeres son feministas</i> - <i>buena foto teodoro pareces una p*ta en decadencia https://t.co/1LecSxbGKi</i>

De manera similar, los casos donde los hombres etiquetan un texto como “sexista” mientras que las mujeres lo consideran “no sexista” suelen estar dirigidos hacia los hombres, expresados de forma sutil o

en tono de broma, lo que tiende a ser más aceptado por el público. Esto refleja la percepción común de que el sexismo se dirige predominantemente hacia las mujeres, aunque en realidad afecta a ambos géneros. En situaciones donde los hombres identifican contenido como sexista y las mujeres no están de acuerdo, generalmente se trata de burlas hacia los hombres, como se muestra en los ejemplos. Estos patrones resaltan las percepciones diversas y matizadas del sexismo en el discurso, lo que subraya la necesidad de abordar estos temas con sensibilidad y comprensión.

Los casos de coincidencia entre mujeres y hombres en el etiquetado son especialmente relevantes porque parecen alinearse con una definición más general y compartida de lo que como sociedad identificamos como sexismo. Cuando tanto mujeres como hombres etiquetan un texto como “sexista”, suelen ser mensajes que contienen estereotipos evidentes, referencias explícitas al acoso o comentarios despectivos que encajan claramente en nociones ampliamente reconocidas de desigualdad de género, como se muestra en la Tabla 6. Por ejemplo, mensajes que destacan la cosificación de las mujeres o hacen burla directa de problemas estructurales del machismo suelen generar un acuerdo universal.

De manera similar, los textos etiquetados como “no sexistas” por ambos grupos también reflejan un consenso social, aunque en un contexto diferente. Estos casos suelen involucrar mensajes neutrales, informativos o humorísticos sin implicaciones de género explícitas, lo que indica que, independientemente de las diferencias individuales, existe un reconocimiento compartido de contenido libre de sexismo.

4.2.2. Percepción del sexismo por edad

Se realizó un análisis enfocado en los grupos de edad de los anotadores, cuyos resultados se presentan en la Tabla 7.

Tabla 7. Diferencias y similitudes entre personas de 18 a 22 años, de 23 a 45 años y de 46 años o más.

		23-45		46+			
		Sexismo	No Sexismo	Sexismo	No Sexismo		
18-22	Sexismo	523	45	23-45	Sexism	530	56
	No Sexismo	63	702		No Sexismo	55	692
		18-22		46+			
		Sexismo	No Sexismo	Sexismo	No Sexismo		
46+	Sexismo	512	73	18-22	Sexismo	512	73
	No Sexismo	56	692		No Sexismo	56	692

Aunque las diferencias entre los grupos etarios —18 a 22 años, 23 a 45 años y 46 años o más— son más sutiles que las observadas por género, también sugieren que las percepciones del sexismo pueden variar en función de la edad. Este enfoque nos permite aportar una visión más completa de cómo las características de los anotadores influyen en la evaluación del sexismo en textos.

Tabla 8. Coincidencia y desacuerdo en las etiquetas de sexismo entre diferentes grupos de edad.

Etiqueta	Tuits
18-22: YES 23-45: NO	<ul style="list-style-type: none"> - @ParentiSol cheto florida 40 te voy a re acosar ahora. - @gonzalezvalen18 Su novio es un bobo hppp. Por eso les pegan cacho. A las mujeres hay que hacerle todo lo que pidan y convertirla. - ¿Qué harán las feminazis ahora que la feminazi máxima, Kena Lorenzini, fue descubierta como acosadora sexual y laboral? Seguro que por ser de izquierda no harán nada. Como ocurrió con Boric - En todo va a meter el feminismo???? Es un debate flaca #DebateCapital
18-22: NO 23-45: YES	<ul style="list-style-type: none"> - @jero_margelich Jajajaa pobre flaca, damela unos días te la devuelvo como una señorita. - Siempre tuvo la frente muy alta, la lengua muy larga y la falda muy corta - @CondeSciendo Ay siiiii, Queen Guarra. Lady poco limpia, ms sucia o la larga vaga. Por qué la votará la gente? Si solo importa la higiene de la casa - @KtaNaldo_Bucara El hacerlo con una mujer embarazada levanta un morbo increíble...Que buen vídeo muchachos.
18-22: YES 46+: NO	<ul style="list-style-type: none"> - @Skywalker6689 Oye, oye, pero ¿qué concepto del amor romántico o de la fidelidad puede tener esa chica?En shock. - No, no, disculpame tú. —Se llevó las manitas a las caderas, relajando los hombros—. Estaba entrenando y siempre se me olvida ducharme antes de hablar con la gente civilizada. —Bromeó—. https://t.co/Pn7kGBU9wl - @BrxnzeGuardian —¿Te he dicho que eres precioso?Preguntó el demonio vampírico con una suave sonrisa en su rostro. Se acercó un poco más, abrazándolo por las caderas. - @PamelaAlemapq @AustinPalao obvio que sii , es guapísimo, chicas un grito saoooooooo!!!
18-22: NO 46+: YES	<ul style="list-style-type: none"> - Aprobación masculi... qué?, sí topas que soy mi fondo de pantalla?, weee soy mi ÍDOLA. Ya mero voy a estar pidiendo aprobación de un hombre cisgenero https://t.co/9LD8TIK1Q1 - @vidadignaYA Es que ahí solo hay manos chicas... ¡Y va por la próstata! - @galvan_1202 yo t dije chacha quien avisa no es traidor - @23_Shephard GORDA P*TA Y TIRA CAUCHO. VOMITIVA.
23-45: YES 46+: NO	<ul style="list-style-type: none"> - @Emmanue50011275 @miky_o La única paloma que me como es la de tu mujer. Gracias por las felicitaciones a mi tribu que definitivamente es más inteligente que tú. Ya deja de avergonzar a tu madre, ella no te crío así - @martupauer @AldairTenorio7 @diana_aceves_ Pero cómo le van a decir a un hombre cómo vestirse? Eso es hembrismo. (Ni hablar del ninguneo al referirse a él despectivamente con la palabra 'tipo' - aunque en muchos contextos se usa sin problemas ni ofensas, éste no parece ser uno de esos) - @alexiarivass @albertdomenech Retratada quedas tu guapi,y nadie te falta el respeto por ser mujer no vengas de feminazi ,te lo decimos como personas me da igual hombre que mujer y eres falsa y muy bajuna y ojalá tu recorrido en TV sea muy corto ,con tu afan de protagonismo eres ridícula, - @LeiiVaquilla Que nai jajaja! Hasta se me aflojaron las caderas de nuevo, que se prepare el 31 y primero jajajaja.
23-45: NO 46+: YES	<ul style="list-style-type: none"> - Pues se acabó. Porque con el vídeo de mañana te cagas en las bragas. https://t.co/k6XNRey6fG - @consumogob @yodominguez ¿Ha dicho .el tamaño importa? Falocentrismo inconsciente - Hacen un doctorado en estudios de género y automáticamente creen que la sociedad les debe un salario de 15 millones - @PerroChusko @MineduPeru @congresoperu Que el de todas el feminismo es cáncer.

Nuestro segundo análisis se centró en los grupos de edad de los anotadores, donde se identificaron

diferencias en la percepción del sexismo, aunque de manera más sutil, entre los grupos de 18 a 22 años, 23 a 45 años y mayores de 46 años, como se ilustra en la Tabla 8.

Como era de esperarse, estos resultados indican una coherencia en las opiniones dentro de ciertos grupos etarios, pero también destacan discrepancias notables en la percepción del sexismo, especialmente al comparar a los participantes más jóvenes (18-22 años) con los mayores de 46 años. Estas discrepancias sugieren que las interpretaciones del sexismo pueden estar influenciadas por contextos generacionales, experiencias personales y cambios socioculturales en las actitudes hacia los roles de género.

Además, el análisis revela que las generaciones más jóvenes tienden a identificar como sexistas ciertos comentarios o actitudes que los participantes de mayor edad consideran neutros o inofensivos. Por otro lado, los mayores de 46 años a menudo perciben como sexistas expresiones que los participantes más jóvenes interpretan como humorísticas o informales. Estas observaciones subrayan la importancia de considerar las dinámicas generacionales al abordar la evaluación y clasificación de discursos relacionados con el sexismo.

4.3. Resumen del capítulo

En este capítulo, se ha analizado el conjunto de datos EXIST, un recurso diseñado para abordar el complejo desafío de identificar y clasificar contenido sexista en redes sociales, considerando las diferencias en la percepción según género y edad. El enfoque metodológico basado en el paradigma de aprendizaje con desacuerdo (*Learning with Disagreement*) ha sido clave para garantizar que las diversas interpretaciones de los anotadores se reflejen en el etiquetado, promoviendo así un análisis más inclusivo y representativo.

Los resultados presentados evidencian la influencia de los factores sociodemográficos en la percepción del sexismo, destacando la importancia de mitigar los sesgos que pueden surgir tanto en la selección de datos como en los procesos de anotación. Este esfuerzo contribuye significativamente al desarrollo de sistemas de PLN más equitativos y sensibles al contexto sociocultural.

Los hallazgos descritos en este capítulo reafirman el compromiso con la construcción de herramientas y recursos lingüísticos que reflejen la diversidad de perspectivas, sentando las bases para futuras investigaciones y aplicaciones que promuevan una interacción más justa y respetuosa en los entornos digitales.

Capítulo 5. Detección Automática de Sexismo mediante Ensamblajes

En este capítulo se presenta una metodología para la detección automática de sexismo en textos, que integra perspectivas de diferentes grupos demográficos mediante modelos de lenguaje natural y técnicas de ensamble. La propuesta emplea un conjunto de datos balanceados y un modelo de clasificación que capta las opiniones de grupos específicos de género y edad, conjuntando estas perspectivas a través del uso de enfoques de ensamble y modelos preentrenados como transformers para tuits, con el apoyo de técnicas de votación por consenso. De esta manera, se logra una representación más inclusiva en la detección de contenido sexista. Para evaluar la eficacia de esta metodología, se utilizó el conjunto de datos EXIST, lo que permite validar la precisión y adaptabilidad del enfoque en un entorno controlado y relevante.

Adicionalmente, aprovechando la gran capacidad de los modelos de lenguaje actuales, como GPT-4o, para generar textos a partir de prompts específicos, se exploró su uso para producir explicaciones detalladas y argumentadas sobre las decisiones del clasificador. Esto no solo facilita la interpretación de los resultados por parte de usuarios humanos, sino que también promueve una mayor transparencia en el modelo y contribuye a la equidad en la detección de contenido sexista, ofreciendo un recurso valioso para la moderación de contenido y la toma de decisiones informadas.

5.1. Preprocesamiento de Datos y Metodología de Clasificación

Dado que el conjunto de datos proviene de textos en Twitter, caracterizados por su naturaleza informal y la inclusión frecuente de elementos irrelevantes como etiquetas o enlaces, se llevó a cabo un preprocesamiento. Las menciones fueron reemplazadas por “@USER” y los enlaces por “HTTPURL”. Este paso buscó eliminar elementos no informativos y optimizar el rendimiento del modelo.

El conjunto de datos fue elaborado con meticulosidad para maximizar la representatividad y la diversidad en la detección de sexismo. En total, contiene 2,526 muestras destinadas al entrenamiento del modelo, 639 muestras adicionales reservadas para la validación y un conjunto final de 490 muestras para una evaluación exhaustiva. Esta división asegura una separación clara de datos entre entrenamiento, validación y prueba, permitiendo evaluar con precisión la efectividad y capacidad de generalización de los modelos propuestos.

Inicialmente, el modelo fue entrenado considerando las opiniones de los seis perfiles por tuit. A partir de un voto mayoritario entre los seis anotadores, se asignó una única etiqueta por tuit. Los resultados de este proceso, representados por las etiquetas obtenidas mediante el voto mayoritario, se ilustran en la Figura 4 (a). Con este enfoque basado en etiquetas consensuadas, nuestro objetivo era evaluar si el modelo podía abarcar múltiples perspectivas reflejadas en las etiquetas obtenidas.

El flujo detallado de esta metodología, desde el preprocesamiento hasta la clasificación final, se presenta en la Figura 3. Este esquema consta de varias etapas de procesamiento y clasificación de textos (tuits) mediante modelos basados en transformers. En este proceso, cada tuit t , extraído del conjunto de datos destinado al entrenamiento del modelo y ya etiquetado de manera binaria (etiqueta dura), pasa primero por una etapa de preprocesamiento. Aquí, las etiquetas son normalizadas y los enlaces en el texto procesados, transformando t en t' .

Luego, el texto preprocesado t' se introduce en un modelo transformer, que clasifica los tuits como sexistas (etiqueta 1) o no sexistas (etiqueta 0). Esta clasificación genera una predicción binaria que se utiliza durante el entrenamiento del modelo. Finalmente, al aplicar esta arquitectura al conjunto de prueba (test), se obtienen las predicciones $p(t')$, que corresponden al resultado de nuestro modelo baseline. En esta etapa, ya es posible determinar si un tuit se clasifica como sexista o no.

5.2. Enfoque Propuesto Basado en Ensamble

Investigaciones previas han demostrado que segmentar a los anotadores en grupos específicos puede mejorar significativamente el rendimiento en tareas de clasificación subjetiva. En esta propuesta, se aplica un enfoque de ensamblado que combina las salidas de clasificadores diseñados para distintos grupos de edad y género. Este método integra las perspectivas únicas de cada grupo demográfico, promoviendo una representación más inclusiva y equitativa. En lugar de depender de una única etiqueta agregada, el modelo captura la diversidad de opiniones individuales, un aspecto clave en tareas como la detección de sexismo.

El modelo propuesto para la identificación de sexismo en tuits se basa en este enfoque de ensamble. Combina las salidas de varios modelos de clasificación individuales, especializados en diferentes grupos de edad y género. A través de un mecanismo de votación mayoritaria, se asigna una etiqueta final cuando al menos cuatro de estos modelos independientes coinciden en la misma categoría.

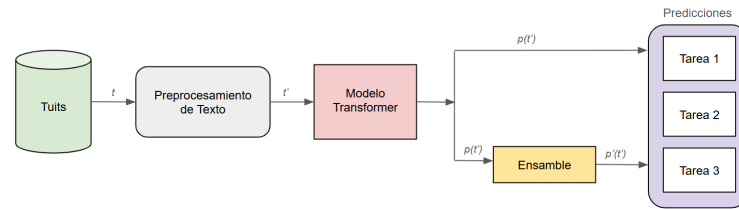


Figura 3. Modelo general para las tres tareas.

La consolidación del conjunto de datos se realizó según género y edad de los anotadores, buscando consenso en las etiquetas. Para las opiniones agrupadas por género, las etiquetas se asignaron si al menos dos de tres anotadores coincidían en clasificar un tuit como sexista o no sexista. En el caso de las opiniones por edad, se utilizaron solo dos anotadores, y las etiquetas se asignaron únicamente si ambos coincidían; de lo contrario, el tuit era descartado. Este enfoque permitió capturar perspectivas específicas de cada grupo demográfico, como género y edad, utilizando un esquema de evaluación rígido.

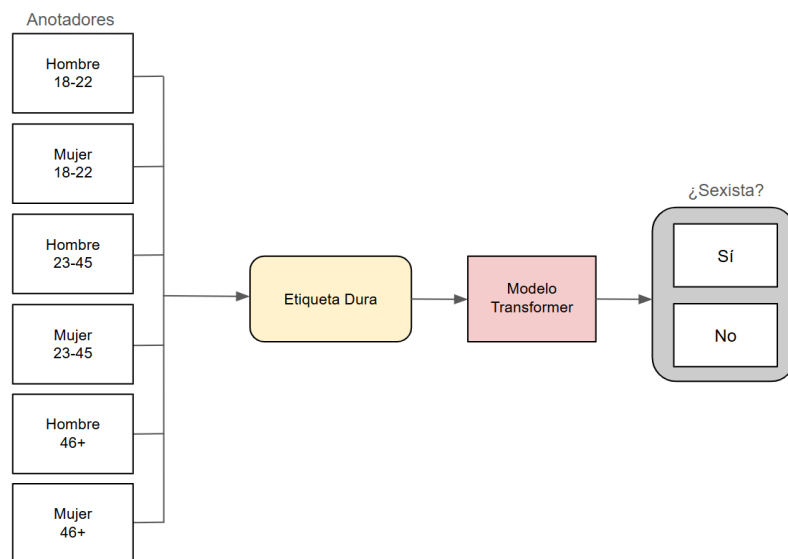
Se establecieron tres grupos de edad: 18-22, 23-45 y 46+. Para las etiquetas basadas en género, se utilizaron únicamente las opiniones de los tres perfiles femeninos o los tres masculinos, aplicando reglas consistentes de consenso. Este proceso asegura un análisis más detallado y representativo de las perspectivas de cada grupo demográfico, permitiendo una clasificación más inclusiva y precisa.

El flujo general del enfoque propuesto se ilustra en la Figura 4 (b). Cada tuit t , extraído del conjunto de datos destinado al entrenamiento del modelo y etiquetado de manera binaria (etiqueta dura), pasa por una etapa inicial de preprocesamiento. Durante esta etapa, se normalizan las etiquetas y se procesan los enlaces contenidos en el texto, transformando t en t' . Este texto preprocesado se introduce luego en modelos transformer independientes que clasifican los tuits como sexistas o no sexistas, generando múltiples predicciones para el mismo tuit $p(t')_1, \dots, p(t')_n$, correspondientes a los distintos modelos del ensamble. Finalmente, mediante un mecanismo de votación mayoritaria, estas predicciones se consolidan en una etiqueta final $p'(t')$, que representa la decisión conjunta del sistema.

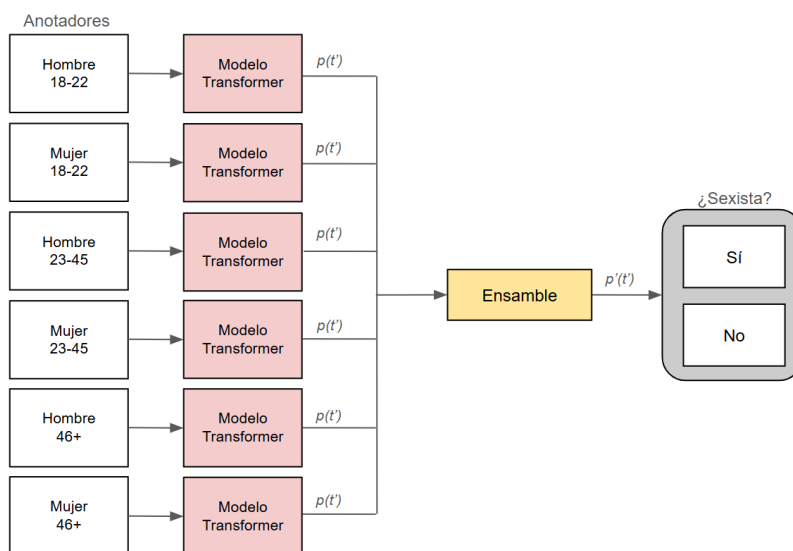
Para garantizar decisiones más precisas, este enfoque incorpora un método de votación dura complementado con un análisis probabilístico. Aunque el mecanismo principal se basa en la elección discreta de cada modelo, se utiliza también la salida softmax del modelo transformer, que asigna una probabilidad a cada categoría, indicando el nivel de confianza en la predicción.

Hemos establecido un umbral de 0.5 para determinar la confiabilidad de una predicción. Por ejemplo, si las probabilidades asociadas a las categorías de un tuit son $[0.50, 0.20, 0.32, 0.43, 0.89]$, evaluamos

cada valor comparándolo con este umbral. Si la probabilidad supera 0.5, se asigna un 1; de lo contrario, un 0. Esto genera una lista binaria, como [1, 0, 0, 0, 1], que indica si el tuit pertenece o no a la categoría "0" (donde la categoría 0 significa que el tuit no es sexista).



(a) Modelo de Etiquetas Duras.



(b) Modelo de Conjunto de Género y Edad para la Identificación de Sexismo con Ensamble.

Figura 4. Modelos para la Identificación de Sexismo

El enfoque de ensamblaje ofrece varios beneficios clave. Primero, mejora la robustez del sistema al integrar perspectivas de modelos entrenados en distintos grupos demográficos. Esto resulta especialmente relevante en problemas sociales como la identificación de sexismo, donde los sesgos y las variaciones

contextuales pueden influir significativamente en las predicciones. Segundo, el modelo optimiza la precisión global al aprovechar la especialización de cada clasificador individual, logrando una adaptación más precisa a las características del conjunto de datos. En última instancia, este método no solo mejora la efectividad en la predicción y clasificación de etiquetas, sino que también asegura una mayor capacidad para manejar la complejidad y la variabilidad inherente a los datos analizados.

5.3. Tarea 2: Intención

Para la identificación de intención de sexismo y categorización, se realizó un análisis enfocado exclusivamente en los tuits previamente identificados como sexistas. Esta decisión buscó profundizar en las características particulares de este tipo de contenido, abordando con mayor precisión los desafíos asociados a su clasificación y comprensión. Al limitar el análisis a este subconjunto, se garantizó que los modelos se concentraran únicamente en el comportamiento lingüístico y las estructuras textuales relacionadas con el sexismo. Esto permitió identificar patrones y matices específicos de este tipo de discurso, reduciendo el ruido en los datos y optimizando la eficiencia del aprendizaje del modelo al centrar los recursos computacionales en los ejemplos más relevantes.

El objetivo de la tarea 2 fue identificar la intención asociada a cada tuit sexista. A diferencia de la tarea anterior, esta problemática presentó mayor complejidad, ya que cada tuit debía asignarse a una única categoría de intención. Para abordar este desafío, se desarrolló un enfoque basado en modelos transformers especializados, cada uno enfocado en un tipo de intención particular.

El proceso comenzó con el preprocesamiento de los tuits, representados como t' , y agrupados según el perfil demográfico correspondiente (por ejemplo, hombres, mujeres, jóvenes de 18-22 años, hombres de 23-45 años, etc.). Cada tuit preprocesado fue evaluado por tres modelos transformers de manera paralela. En el modelo correspondiente a la intención específica del tuit, este fue etiquetado como sexista (valor 1). En los otros dos modelos, donde la intención no correspondía, el tuit fue etiquetado como no sexista (valor 0). Este etiquetado binario optimizó el entrenamiento de los modelos, permitiéndoles distinguir entre una categoría específica de intención y las demás.

Una vez completado este proceso, los modelos generaron predicciones para los tuits preprocesados ($p(t')$). En el paso de ensamble, se seleccionó la categoría con la probabilidad más alta para cada tuit, asignándole así la intención más probable. Posteriormente, estas predicciones fueron refinadas mediante un esquema

de votación dura basado en probabilidades, consolidando las etiquetas finales. Este flujo asegura que cada tuit sea clasificado únicamente en la categoría de intención más relevante.

El enfoque completo se detalla en la Figura 5, donde se observa cómo un tuit pasa por los modelos correspondientes, es procesado y obtiene una categoría de intención única a través del ensamble. Este procedimiento permitió abordar de manera eficiente las complejidades inherentes a esta tarea, maximizando la precisión de las predicciones.

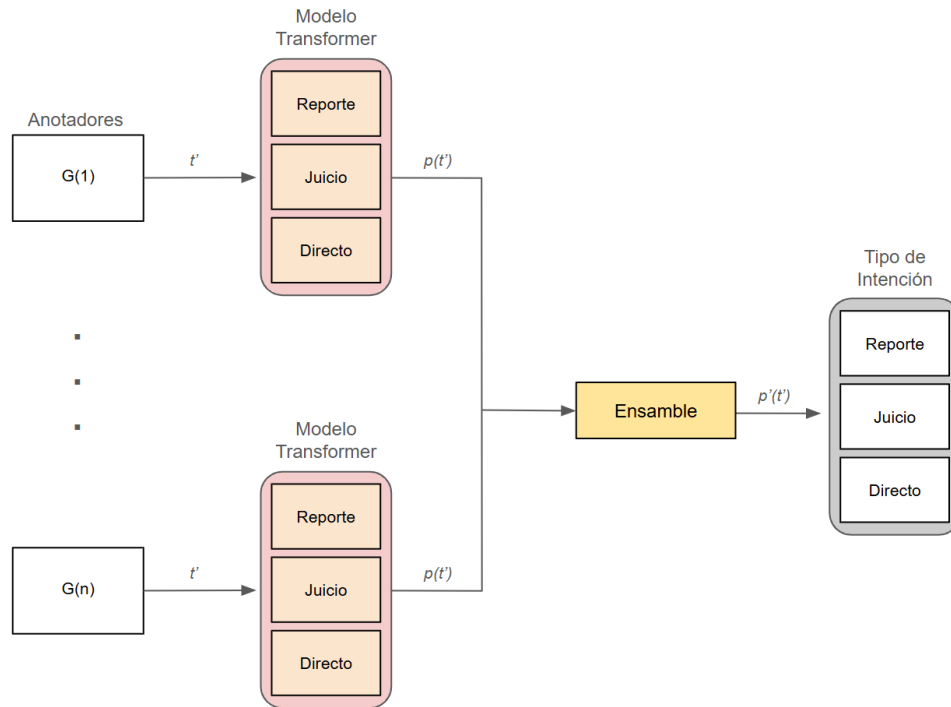


Figura 5. Modelo para la Detección de Intención.

5.4. Tarea 3: Categorización

En la tercera tarea, el enfoque se centró en la categorización de los tuits sexistas, permitiendo que un mismo tuit pudiera pertenecer a múltiples categorías. Por ejemplo, un tuit podía clasificarse simultáneamente como misoginia y violencia sexual. Esta flexibilidad incrementó la complejidad del problema, ya que las etiquetas debían reflejar todas las categorías aplicables a cada texto de forma precisa.

Para abordar este desafío, se conservaron todas las categorías asignadas por los anotadores al momento de generar las etiquetas duras. Cada tuit preprocesado (t') fue procesado por cinco modelos transformers

independientes, cada uno especializado en una categoría específica. Si un tuit pertenecía a una categoría particular, fue etiquetado como sexista para ese modelo; de lo contrario, fue etiquetado como no sexista. Por ejemplo, un tuit etiquetado como misoginia y violencia sexual fue considerado sexista en los modelos correspondientes a estas categorías, mientras que en los tres restantes fue tratado como no sexista.

Tras el procesamiento, se obtuvieron las predicciones $p(t')$ para el conjunto de prueba. Estas predicciones se combinaron mediante un ensamble basado en votación dura con enfoque probabilístico, considerando todas las categorías asignadas por los modelos. El resultado final, $p'(t')$, representa las predicciones ajustadas tras el proceso de ensamble, asignando a cada tuit las categorías correspondientes, por ejemplo: ["JUICIO", "DIRECTO", "VIOLENCIA-SEXUAL"].

Este enfoque permitió capturar las complejas relaciones entre las categorías de sexismo, generando resultados más completos y representativos. La metodología aplicada se ilustra en la Figura 6, donde se muestra cómo un tuit pasa por los modelos especializados, se procesan sus predicciones y se consolidan las etiquetas finales a través del ensamble.

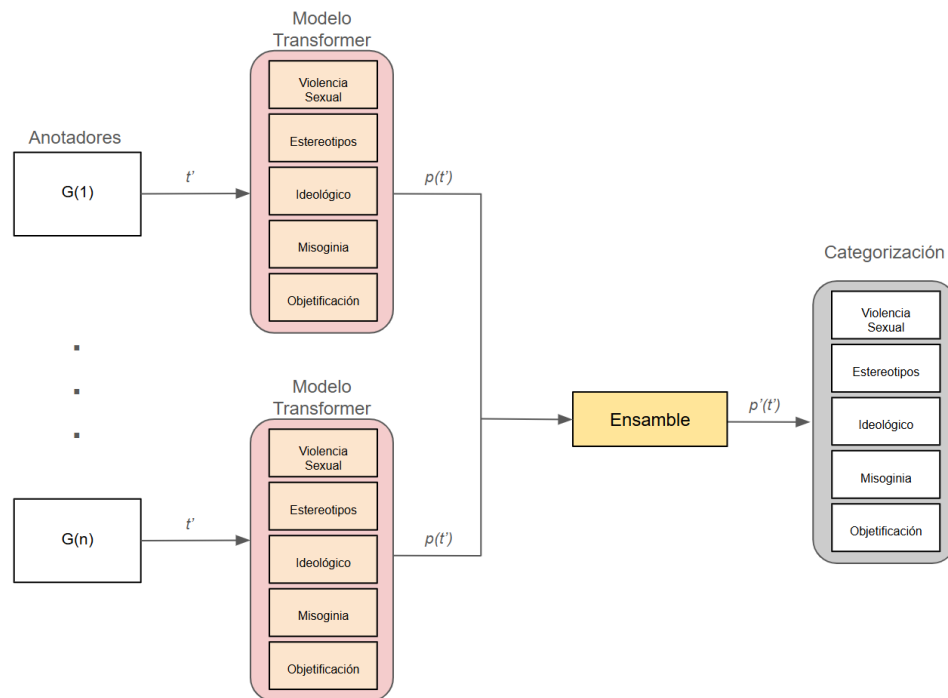


Figura 6. Modelo para la Categorización del Sexismo.

5.5. Explicación de las predicciones del modelo con ChatGPT

Con el objetivo de ofrecer una justificación explícita y coherente para cada etiqueta asignada, se busca incrementar la confianza en las decisiones automatizadas generadas por el modelo GPT-4o (2024-08-01).

Este enfoque detalla de manera explícita las razones detrás de estas clasificaciones. La integración de explicaciones detalladas mejora la transparencia del sistema, reduciendo el riesgo de malinterpretaciones y permitiendo un análisis más profundo del contenido. La metodología empleada implica el uso de un prompt cuidadosamente diseñado, que guía a GPT-4o para que analice y explique cada decisión del modelo. En la Figura 7 se muestra el prompt utilizado.

Al proporcionar explicaciones generadas por un modelo de lenguaje, se logra una mayor transparencia al acompañar cada decisión con una justificación basada en el contenido y contexto del texto. Esto también contribuye a reducir malinterpretaciones, ya que las explicaciones ayudan a comprender mejor los elementos clave del análisis. Además, estas interpretaciones detalladas son de gran utilidad para moderadores humanos, quienes pueden revisar las justificaciones para tomar decisiones informadas en la gestión de contenido.

Este enfoque permite que las decisiones del modelo se adapten al contexto, considerando las sutilezas del lenguaje y las diferencias culturales, lo que mejora significativamente la precisión y relevancia de las predicciones. Al combinar predicciones y explicaciones detalladas, se fomenta un sistema de detección de sexismo que no solo es preciso, sino también comprensible y aplicable en entornos reales de moderación de contenido. Este proceso promueve un mayor grado de confianza en el modelo y dota a los moderadores de herramientas sólidas para realizar su labor de manera efectiva.

Hola,

Necesito ayuda para explicar por qué estos tuits en español fueron clasificados como sexistas o no sexistas por mi modelo, que fue entrenado con opiniones de diferentes personas en diferentes rangos de edad y género, por lo que cada etiqueta es su consenso. El objetivo es una clasificación binaria, donde 1 indica que un tuit es sexista y 0 indica que no lo es. El sistema debe determinar si un tuit contiene expresiones o comportamientos sexistas, es decir, si es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista. A continuación, se muestran ejemplos de tuits clasificados como sexistas y no sexistas, junto con una breve explicación de cada caso. Es importante destacar que, independientemente de si los tuits fueron clasificados como sexistas o no sexistas, necesito una explicación de la razón de su clasificación para cada tuit.

Tuits Sexistas (1):

- "¡Mujer al volante, ten cuidado!"

Explicación: Este tuit perpetúa un estereotipo negativo sobre las mujeres, sugiriendo que son malas conductoras. Esto es un ejemplo claro de sexismo porque hace una generalización negativa basada en el género.

"La gente realmente intenta convencer a las mujeres con poco o ningún trasero de que deberían salir y comprarse un cuerpo. Como bih, no necesito un trasero gordo para conseguir un hombre. Nunca tiene."

Explicación: Este tuit refuerza estereotipos sobre el cuerpo femenino y sugiere que la valía de una mujer está ligada a su apariencia física, lo que es una forma de sexismo. También reduce a las mujeres a meros objetos de deseo.

Tuits No Sexistas (0):

- "Alguien me explica que zorra hace la gente en el cajero que se demora tanto."

Explicación: Aunque el lenguaje es vulgar y ofensivo, no está dirigido específicamente a las mujeres ni perpetúa estereotipos de género. La crítica se dirige a la impaciencia en una situación común, sin connotaciones sexistas claras.

- "@messyworldorder, sinceramente, es muy vergonzoso verlo y dirán: 'no todas las mujeres blancas son así'"

Explicación: Este tuit parece estar comentando sobre un comportamiento o una percepción sin atacar directamente a las mujeres en general. La frase "no todas las mujeres blancas son así" parece referirse a una crítica a una percepción específica sin implicar un juicio sexista.

En resumen, la clasificación se basa en si el tuit refuerza estereotipos de género, desvaloriza a las mujeres o perpetúa ideas sexistas. Independientemente de si los tuits fueron clasificados como sexistas o no sexistas, necesito una explicación para cada tuit sobre cómo fue clasificado. Espero que esta explicación sea útil para comprender cómo funciona el modelo de clasificación. Estos son los tuits:

@anacaotica88 @MordorLivin No me acuerdo de los detalles de GamerGate, pero ella estuvo en el ojo del huracán recibiendo acoso de hombres indignados (y sus medios frikis) y creo que también acosaron a Brianna Wu, q es transfemenino. Seguramente tuvo eso que ver y quiso cerrar filas". 1

@cosmicJunkBot lo digo cada pocos días y lo repito: TODO esto se pudo evitar si se hubiera parado el gamergate a tiempo 0

Figura 7. Prompt utilizado.

5.6. Resumen del capítulo

Este capítulo presenta una metodología que combina técnicas avanzadas de PLN, como los modelos de transformación y ensamble, para abordar de manera inclusiva y precisa la detección de sexismo en textos. Se han expuesto las ventajas de utilizar un enfoque de ensamble que respeta las particularidades de distintos grupos demográficos, así como la importancia de realizar un análisis de errores detallado para mejorar la precisión y reducir los sesgos del modelo. Al incorporar explicaciones generadas con GPT-4, se aporta transparencia a las predicciones del modelo, facilitando su aplicación en escenarios reales de moderación de contenido. En conjunto, esta propuesta busca no solo mejorar la precisión de la detección de sexismo, sino también ofrecer un enfoque interpretativo y adaptado a la diversidad demográfica, contribuyendo a la construcción de herramientas de moderación más justas y efectivas.

Capítulo 6. Resultados

Este capítulo presenta los resultados sobre la identificación automática del sexismo en redes sociales, centrándonos particularmente en cómo las percepciones de los anotadores, diferenciadas por género y edad, influyen en los resultados de los modelos de clasificación. Para llevar a cabo un análisis de error más profundo, primero utilizamos redes de atención jerárquicas, que nos permitieron identificar con precisión las áreas de mejora del sistema al examinar casos de falsos positivos y falsos negativos. Posteriormente, empleamos ChatGPT para complementar este análisis, proporcionando explicaciones sobre las decisiones tomadas por el modelo y obteniendo una comprensión más detallada de los resultados y las posibles causas de los errores.

6.1. Preprocesamiento de los datos

En este estudio se comparó el desempeño del modelo utilizando tres tipos de datos: crudos, semi limpios y limpios, a fin de determinar cuál ofrecía un mejor equilibrio entre simplicidad y efectividad.

En primer lugar, los datos crudos se mantuvieron tal cual se encontraban en el conjunto de datos original, sin aplicar ningún tipo de limpieza. En segundo lugar, los datos semi limpios fueron preprocesados reemplazando las menciones (@) por @USER y los enlaces (<http://> o <https://>) por HTTPURL; este paso tenía como objetivo mejorar el rendimiento del modelo al eliminar elementos no informativos del texto. Por último, los datos limpios fueron sometidos a un proceso de limpieza más exhaustivo, eliminando menciones, enlaces, hashtags, números, contenido entre paréntesis y corchetes, saltos de línea, signos de puntuación como puntos, comas y signos de exclamación o interrogación, así como espacios extra.

Los ejemplos de cada uno de estos casos se presentan a continuación:

- **Crudo:** @andreamh... Las zorras del zodiaco son tan ridlculas <https://t.co/5eooVmu6lv>
- **Semi Limpio:** @USER Las zorras del zodiaco son tan ridlculas HTTPURL
- **Limpio:** las zorras del zodiaco son tan ridiculas

En términos de rendimiento, se evaluaron los resultados mediante el F1 score, utilizando la metodología de etiquetas duras ilustrada en la Figura 4 (a) del Capítulo 5. Los experimentos fueron realizados con

el modelo de ensamble descrito en la Sección 5.2 (Capítulo 5). Los valores promedio obtenidos fueron los siguientes: 0.8319 para los datos crudos, 0.8385 para los datos semi limpios y 0.8235 para los datos limpios. Como se ilustra en la Figura 8, el modelo alcanzó su mejor desempeño utilizando los datos semi limpios.

Con base en estos resultados, se puede concluir que, aunque la limpieza completa del texto ofrece un procesamiento más riguroso, también puede eliminar información relevante para el modelo, como los hashtags, los cuales podrían contener palabras importantes para su desempeño. Por otro lado, el uso de datos crudos incluye ruido que afecta ligeramente el desempeño del modelo. Los datos semi limpios, al equilibrar la reducción de ruido y la preservación de patrones relevantes, demostraron ser la mejor opción en términos de rendimiento, por lo que se decidió utilizar esta versión para los análisis posteriores.

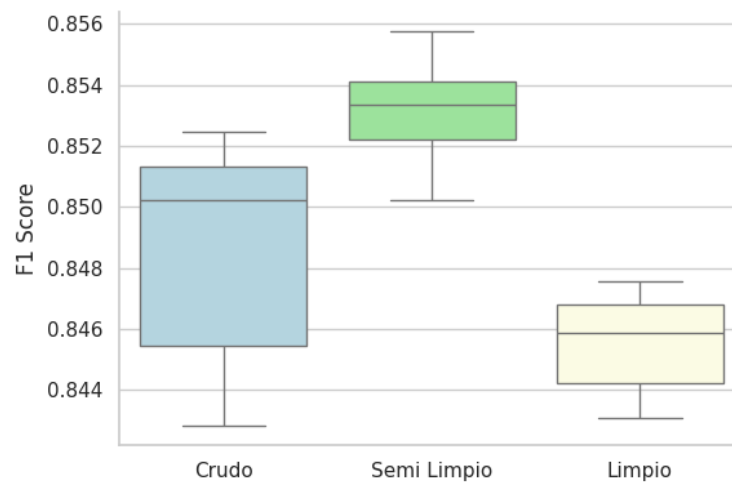


Figura 8. Comparación del rendimiento del modelo al predecir con diferentes tipos de datos.

6.2. Identificación Automática del Sexismo

Comenzamos identificando el sexismo mediante el consenso alcanzado a partir de las seis etiquetas asignadas por tuit. Consideramos una etiqueta como válida (“etiqueta dura”) si al menos cuatro de los seis anotadores coincidían en su clasificación. Este criterio de consenso se utilizó como nuestra línea base (baseline) para comparar los resultados de los experimentos posteriores.

Para realizar este análisis, se utilizaron 11 modelos distintos, como se detalla en la Figura 3(c) del Capítulo 5. Estos modelos consideraron el género (hombres y mujeres), los rangos de edad (18-22, 23-45 y 46+), y combinaciones específicas de género y rango de edad (mujeres y hombres dentro de cada uno

de los tres rangos). Cada modelo fue entrenado para capturar las perspectivas específicas de su grupo correspondiente. Esto implicó la creación de modelos independientes, como uno exclusivo para mujeres, otro para personas de 18 a 22 años, y así sucesivamente para cada uno de los grupos definidos. En todos los casos, las predicciones de los modelos se compararon con las “etiquetas duras” proporcionadas por los seis anotadores, lo que permitió evaluar su desempeño de manera consistente.

Nuestro modelo fue entrenado utilizando tres enfoques diferentes: considerando las opiniones de los seis perfiles, los perfiles femeninos, masculinos y por edades (Fig. 9).

Identificación del sexismo

Observando la Figura 9 (a) se obtuvieron de F1 score para la etiqueta dura (baseline) 0.829 ($\sigma = 0.007$), hombres de 18 a 22; 0.811 ($\sigma = 0.015$), hombres de 23 a 45; 0.808 ($\sigma = 0.008$), hombres de 46 en adelante; 0.779 ($\sigma = 0.011$), mujeres de 18 a 22; 0.799 ($\sigma = 0.008$), mujeres de 23 a 45; 0.824 ($\sigma = 0.010$), mujeres de 46 en adelante; 0.815 ($\sigma = 0.009$), ensamble de mujeres y hombres entre todas las edades; 0.829 ($\sigma = 0.006$).

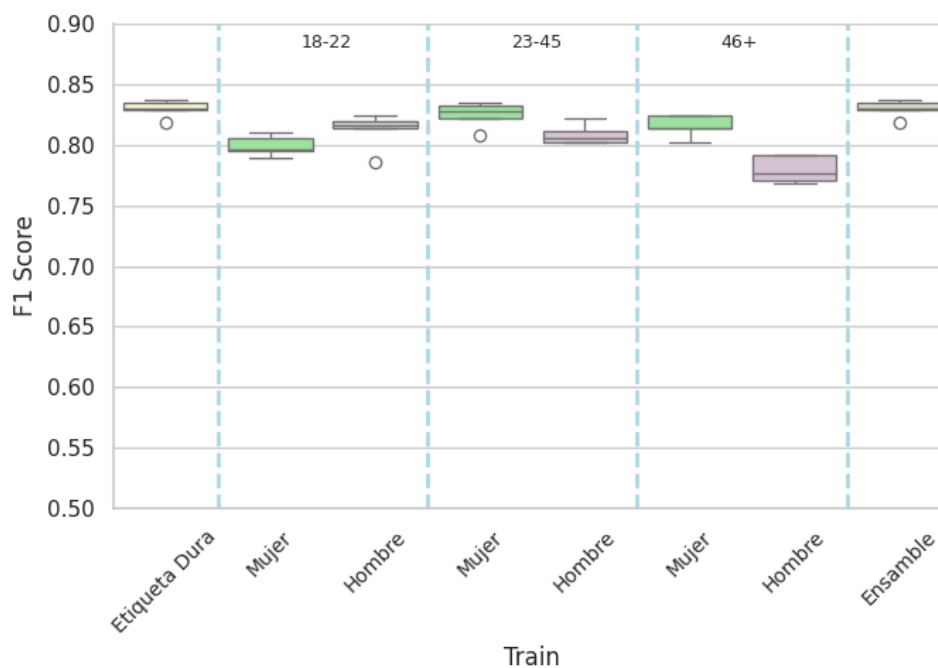
Un análisis de varianza (ANOVA) identificó diferencias significativas entre los grupos analizados ($p < 0.001$), justificando un análisis post hoc de Tukey para determinar las comparaciones específicas.

El análisis de Tukey reveló que no hay diferencias significativas entre los grupos, pero el grupo de ensamble mostró diferencias significativas con varios subgrupos. En particular, hombres de 46+ tuvo valores promedio significativamente más bajos que el ensamble ($p < 0.001$), y los grupos de mujeres de 18 a 22 y hombres de 23 a 45 también presentaron diferencias significativas ($p < 0.05$). Estas diferencias resaltan que el desempeño del ensamble está consistentemente alineado con valores promedio más equilibrados, en contraste con las variaciones observadas en estos subgrupos.

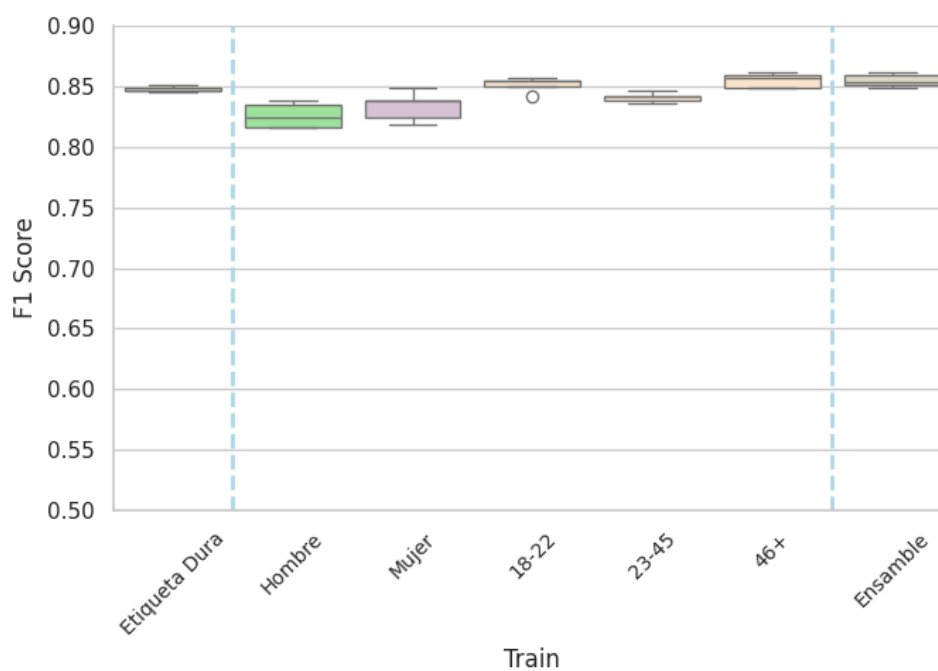
Por otro lado, no se encontraron diferencias significativas entre el ensamble y los grupos mujeres de 23 a 45 ni mujeres de 46 en adelante ($p > 0.05$), lo que indica una mayor similitud en los valores promedio entre estos grupos. Este resultado posiciona al ensamble como un grupo con un desempeño representativo y equilibrado, particularmente en relación con grupos de mujeres en diferentes rangos de edad.

Las comparaciones entre mujeres de 23 a 45 tuvo valores promedio significativamente más altos que hombres de 46 en adelante ($p < 0.001$), y también se identificaron diferencias significativas entre mujeres de 23 a 45 y mujeres de 18 a 22 ($p < 0.01$). Sin embargo, no se encontraron diferencias significativas entre hombres de 18 a 22 y hombres de 23 a 45 ($p > 0.05$), ni entre mujeres de 46+ y mujeres de 23 a

45 ($p > 0.05$).



(a) Comparación según el Perfil de Anotador y su Ensemble.



(b) Comparación entre Diferentes Grupos de Edad y Género, y su Ensemble.

Figura 9. Identificación del Sexismo (Tarea 1).

El grupo ensemble destaca como un punto de referencia clave en el análisis, mostrando un desempeño consistente y comparable con ciertos subgrupos mientras se diferencia significativamente de otros. Este

hallazgo subraya la robustez y representatividad del ensamble en el contexto de los valores promedio observados, especialmente al compararlo con subgrupos masculinos de diferentes rangos de edad.

Observando la Figura 9 (b) se obtuvieron de F1 score para la etiqueta dura (baseline) 0.847 ($\sigma = 0.002$), mujeres y hombres de 18 a 22; 0.851 ($\sigma = 0.005$), mujeres y hombres de 23 a 45; 0.840 ($\sigma = 0.003$), mujeres y hombres de 46 en adelante; 0.855 ($\sigma = 0.005$), hombres; 0.826 ($\sigma = 0.010$), mujeres; 0.833 ($\sigma = 0.012$), ensamble de mujeres, hombres, personas de 18-22, 23-45 y 46+; 0.854 ($\sigma = 0.005$).

Se encontraron diferencias significativas entre los grupos analizados ($p < 0.001$), lo que indica que al menos uno de los grupos posee una media que difiere significativamente de los demás.

El análisis mostró diferencias significativas entre varios subgrupos. En particular, los valores promedio del grupo hombres fueron significativamente más bajos que los del grupo 18-22 ($p < 0.05$), y los del grupo de mujeres también resultaron más bajos respecto al grupo de 18 a 22 ($p < 0.05$). Además, se observaron diferencias significativas entre el grupo de hombres y otros grupos, como 46+, ensamble y la etiqueta dura ($p < 0.05$). De manera similar, el grupo de mujeres presentó valores significativamente más bajos que el grupo de 46+, ensamble y la etiqueta dura ($p < 0.05$).

Por otro lado, algunas comparaciones entre subgrupos no revelaron diferencias significativas. Por ejemplo, no se encontraron diferencias significativas entre los grupos de 23 a 45 y 18 a 22 ($p > 0.05$), ni entre el ensamble y los de 18 a 22 ($p > 0.05$).

Se destacan diferencias significativas entre algunos subgrupos, especialmente aquellos relacionados con el género y los rangos de edad.

Intención de la fuente

En cuanto a la detección de la intención sexista, de la Figura 10 (a) se obtuvieron valores de F1 score para la etiqueta dura (baseline) 0.583 ($\sigma = 0.007$), hombres de 18 a 22; 0.589 ($\sigma = 0.006$), hombres de 23 a 45; 0.547 ($\sigma = 0.012$), hombres de 46 en adelante; 0.575 ($\sigma = 0.010$), mujeres de 18 a 22; 0.658 ($\sigma = 0.014$), mujeres de 23 a 45; 0.617 ($\sigma = 0.010$), mujeres de 46 en adelante; 0.534 ($\sigma = 0.014$), ensamble de mujeres y hombres entre todas las edades; 0.636 ($\sigma = 0.011$).

En cuanto a la detección de la intención sexista, el análisis de varianza (ANOVA) identificó diferencias significativas entre los grupos ($p < 0.001$), lo que permitió realizar un análisis post hoc de Tukey para explorar comparaciones específicas entre los grupos.

El análisis de Tukey mostró que el ensamble presentó diferencias significativas con varios otros grupos.

En particular, los grupos hombres de 23 a 45, hombres de 46+, y mujeres de 46+ tuvieron valores promedio considerablemente más bajos que el ensamble ($p < 0.001$ en todos los casos). Asimismo, el grupo de hombres de 18-22 también mostró valores significativamente inferiores ($p < 0.001$).

Por otro lado, no se encontraron diferencias significativas entre el ensamble y el grupo mujeres de 23 a 45 ($p > 0.05$), lo que indica una similitud en los valores promedio entre estos grupos. Este hallazgo refuerza la representatividad del ensamble como un punto de referencia para los valores promedio observados en este análisis.

Las diferencias entre mujeres de 23 a 45 y mujeres de 18 a 22 también fueron significativas ($p < 0.001$), mientras que las mujeres de 46+ mostró valores significativamente más bajos ($p < 0.001$ en ambos casos).

Estos resultados subrayan la utilidad del ensamble como un punto de equilibrio en el conjunto de grupos analizados.

Observando la Figura 10 (b) se obtuvieron de F1 score para la etiqueta dura (baseline) 0.583 ($\sigma = 0.007$), mujeres y hombres de 18 a 22; 0.617 ($\sigma = 0.007$), mujeres y hombres de 23 a 45; 0.607 ($\sigma = 0.007$), mujeres y hombres de 46 en adelante; 0.591 ($\sigma = 0.004$), hombres; 0.570 ($\sigma = 0.006$), mujeres; 0.615 ($\sigma = 0.002$), ensamble de mujeres, hombres, personas de 18-22, 23-45 y 46+; 0.604 ($\sigma = 0.002$).

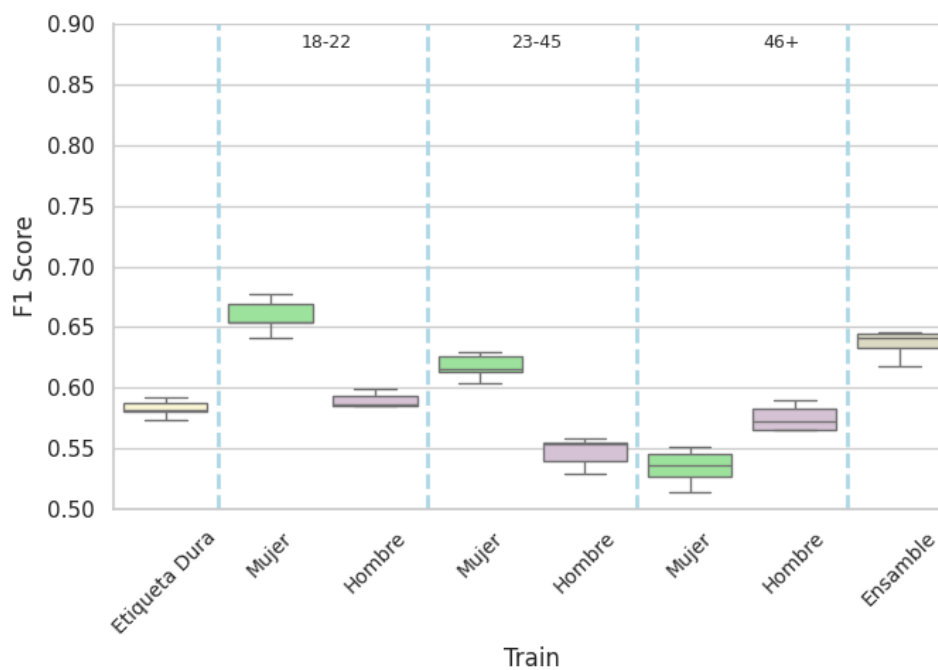
Los resultados indicaron que el ensamble presentó valores promedio significativamente más altos que los grupos hombres, la etiqueta dura y las personas de 46+ ($p < 0.001$ en todos los casos). Sin embargo, no se encontraron diferencias significativas entre el ensamble y el grupo mujeres ($p > 0.05$), lo que sugiere una similitud notable entre estos dos grupos.

Por otro lado, al comparar los grupos por rango de edad, se observaron diferencias relevantes. El grupo 18-22 mostró valores significativamente más altos que 23-45, 46+ ($p < 0.001$ en todos los casos). Asimismo, las personas de 23-45 presentaron valores significativamente superiores a los de 46+ ($p < 0.001$), lo que resalta una tendencia decreciente en los valores promedio conforme aumenta la edad.

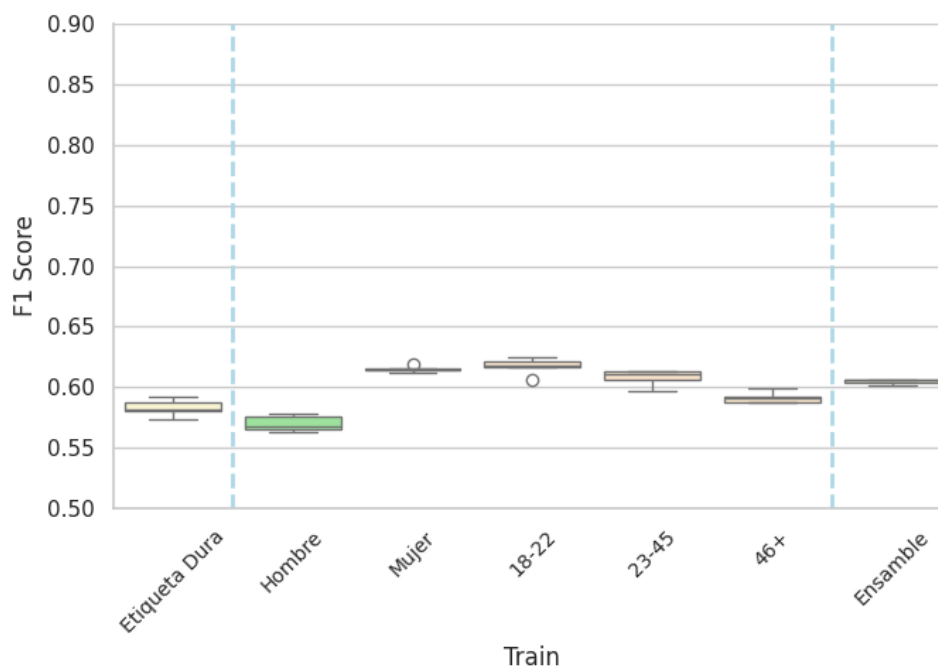
Además, las diferencias entre hombres y los demás grupos fueron notables, con hombres consistentemente mostrando valores más bajos en comparación con mujeres y ensamble ($p < 0.001$). Por su parte, la etiqueta dura también exhibió valores promedio significativamente inferiores a los del ensamble y las mujeres ($p < 0.001$).

En conclusión, el ensamble refleja una proximidad significativa con ciertos grupos, como mujeres, y

diferenciándose de otros, como hombres y la etiqueta dura.



(a) Comparación según el Perfil de Anotador y su Ensemble.



(b) Comparación entre Diferentes Grupos de Edad y Género, y su Ensemble.

Figura 10. Detección de la Intención (Tarea 2).

Categorización del sexismo

Por último, en la categorización del sexismo, de la Figura11 (a) se obtuvieron de F1 score para la etiqueta

dura (baseline) 0.647 ($\sigma = 0.001$), hombres de 18 a 22; 0.613 ($\sigma = 0.004$), hombres de 23 a 45; 0.611 ($\sigma = 0.003$), hombres de 46 en adelante; 0.600 ($\sigma = 0.002$), mujeres de 18 a 22; 0.601 ($\sigma = 0.001$), mujeres de 23 a 45; 0.593 ($\sigma = 0.005$), mujeres de 46 en adelante; 0.600 ($\sigma = 0.001$), ensamble de mujeres y hombres entre todas las edades; 0.699 ($\sigma = 0.0007$).

El grupo ensamble presentó valores promedio significativamente superiores a todos los demás grupos ($p < 0.001$ en todas las comparaciones). Específicamente, las diferencias con los grupos de hombres de 18-22, hombres de 23-45, hombres de 46+, mujeres de 18-22, mujeres de 23-45, y mujeres de 46+ fueron sustanciales, subrayando la posición del ensamble como el punto de referencia más alto.

Dentro de las comparaciones entre grupos sin incluir al ensamble, se identificaron patrones relevantes. Por ejemplo, no se encontraron diferencias significativas entre hombres de 18-22 y hombres de 23-45 ($p > 0.05$), pero los valores de hombres de 46+ fueron significativamente menores en comparación con los hombres de 18-22 y hombres de 23-45 ($p < 0.001$). Similarmente, mujeres de 18-22 y mujeres de 46+ no mostraron diferencias significativas entre sí ($p > 0.05$), pero las mujeres de 23-45 tuvieron valores promedio significativamente menores que las mujeres de 18-22 ($p < 0.01$).

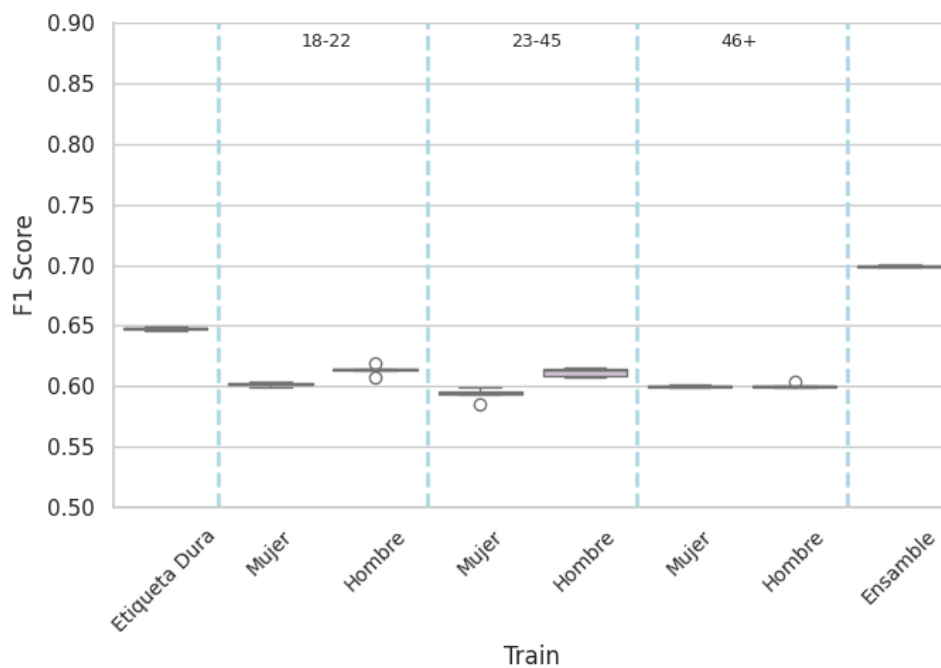
Observando la Figura 11 (b) se obtuvieron de F1 score para la etiqueta dura (baseline) 0.647 ($\sigma = 0.001$), mujeres y hombres de 18 a 22; 0.633 ($\sigma = 0.001$), mujeres y hombres de 23 a 45; 0.635 ($\sigma = 0.003$), mujeres y hombres de 46 en adelante; 0.627 ($\sigma = 0.001$), hombres; 0.668 ($\sigma = 0.002$), mujeres; 0.588 ($\sigma = 0.002$), ensamble de mujeres, hombres, personas de 18-22, 23-45 y 46+; 0.717 ($\sigma = 0.0008$).

El análisis mostró que el ensamble presentó diferencia significativa con otros grupos. Los valores promedio del grupo fueron significativamente mayores en comparación con los otros grupos, incluyendo hombres de 18-22, mujeres de 18-22, hombres de 23-45, mujeres de 23-45, hombres de 46+, mujeres-46+, y con la etiqueta dura (baseline), con diferencias observadas en todos los casos ($p < 0.001$).

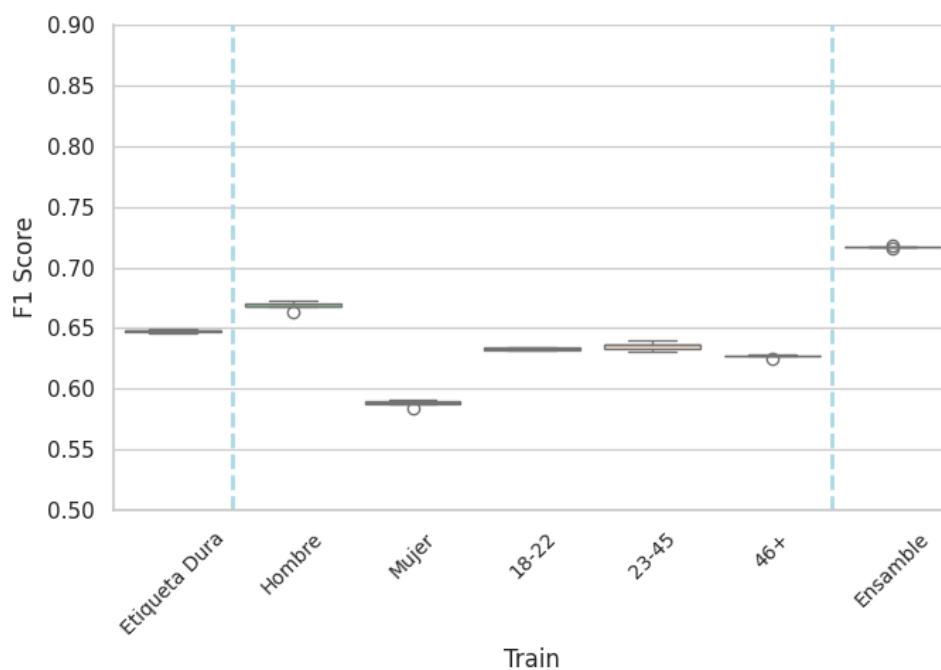
Particularmente, las diferencias entre el ensamble y mujeres ($p < 0.001$) fueron muy notables, mostrando valores significativamente más bajos en comparación con el ensamble.

En cuanto a los grupos de edades, el análisis mostró que las diferencias entre los grupos de 18-22 y 23-45 no fueron significativas ($p > 0.05$), pero las personas de 46+, tuvieron un desempeño significativamente menor ($p < 0.001$) en comparación con lo de 18-22 y 23-45.

Además, el grupo *mujeres* mostró un rendimiento significativamente inferior a los de los hombres en todas las comparaciones realizadas ($p < 0.001$).



(a) Comparación según el Perfil de Anotador y su Ensemble.



(b) Comparación entre Diferentes Grupos de Edad y Género, y su Ensemble.

Figura 11. Categorización del Sexismo (Tarea 3).

El análisis revela que el grupo de ensemble tiene un rendimiento significativamente superior al de los demás grupos, mientras que las comparaciones entre los grupos de diferentes rangos de edad y género refuerzan la importancia de la edad en la variación observada en los valores promedio.

6.3. Análisis de Error

Para comprender en profundidad el desempeño de nuestro modelo y diagnosticar áreas de mejora, realizamos un análisis exhaustivo de los errores cometidos por el modelo de ensamble. Este análisis es esencial para identificar patrones de fallos, optimizar la capacidad del modelo para diferenciar entre contenido sexista y no sexista, y reducir posibles sesgos inherentes a las predicciones.

El análisis se llevó a cabo utilizando un enfoque basado en la interpretación de las decisiones del modelo, que incluye la evaluación de la importancia asignada a diferentes palabras y frases dentro de los textos analizados. Para ello, se empleó un mecanismo de atención que calcula puntuaciones asociadas a cada palabra, las cuales son posteriormente normalizadas mediante la función softmax (Yang et al., 2016). Este procedimiento no solo facilita una interpretación más detallada del comportamiento interno del modelo, sino que también permite detectar qué características lingüísticas tienen mayor peso en las predicciones, revelando posibles puntos débiles o sesgos en el proceso de clasificación.

En nuestro enfoque, incorporamos las Redes de Atención Jerárquica (HAN) descritas en la sección 2.2.2 para analizar de manera segmentada el desempeño en subgrupos específicos, tales como mujeres, hombres y distintas categorías de edad (18-22, 23-45 y 46+). Este análisis granular permitió explorar cómo las diferencias demográficas afectan la distribución de las puntuaciones de atención y las predicciones del modelo. Cada subgrupo generó un conjunto de puntuaciones de atención que destacan las palabras y expresiones más influyentes en la toma de decisiones del modelo para ese grupo en particular.

A partir de estos análisis individuales, consolidamos los resultados en un modelo de ensamble, combinando los valores de atención de los subgrupos mediante técnicas de agregación. Este modelo de ensamble no solo mejora la robustez general del sistema, sino que también proporciona una visión más amplia del rendimiento al unificar las perspectivas de los diferentes segmentos.

El análisis de error se enfocó en identificar tendencias recurrentes, como la presencia de falsos positivos y falsos negativos, evaluando cómo estas variaciones se distribuyen entre los subgrupos. Por ejemplo, observamos que ciertas expresiones o palabras podían ser clasificadas incorrectamente como sexistas en un subgrupo, mientras que en otros no generaban el mismo efecto. Este tipo de información es crucial para refinar el modelo, ya que apunta directamente a los sesgos o limitaciones que podrían estar afectando su capacidad de generalización.

Para profundizar en el análisis de errores, examinamos primero la matriz de confusión de la Tabla 9,

que proporciona una visión detallada del rendimiento de nuestro clasificador al distinguir entre contenido sexista y no sexista (Tarea 1).

Tabla 9. Matriz de Confusión de Predicciones.

		Etiquetas Verdaderas	
		Sexismo	No Sexismo
Etiquetas	Sexismo	244	155
	No Sexismo	17	74

La matriz de confusión nos muestra las cantidades de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. En este caso, observamos que el modelo identificó correctamente 244 casos de sexismo, pero clasificó incorrectamente 155 instancias de no sexismo como sexismo (falsos positivos). Además, hubo 17 casos de sexismo que el modelo no detectó (falsos negativos) y 74 instancias correctamente identificadas como no sexistas.

Este análisis revela que el modelo presenta una mayor cantidad de falsos positivos que falsos negativos, lo que indica una tendencia a clasificar en exceso como sexista el contenido no sexista. Este patrón sugiere que el modelo puede estar reaccionando de manera sensible a ciertas palabras o frases que, si bien suelen aparecer en contextos sexistas, no necesariamente representan contenido sexista en algunos casos. Las puntuaciones de atención obtenidas de los subgrupos demográficos pueden ayudar a identificar estas palabras o frases específicas que generan confusión en el modelo. Por ejemplo, términos o expresiones comunes en algunos grupos de edad o género podrían estar llevando a estas clasificaciones incorrectas.

Los resultados indican una tasa de falsos positivos del 38.8 % y una tasa de falsos negativos del 18.6 %. Los falsos positivos, en los cuales el modelo etiqueta erróneamente contenido no sexista como sexista, pueden llevar a una censura injustificada o alienación de discursos inofensivos. Por el contrario, los falsos negativos, donde el modelo no logra identificar contenido sexista, presentan el riesgo de permitir que discursos dañinos queden sin abordar. Estos errores destacan la importancia crítica de afinar el modelo para equilibrar la sensibilidad y la especificidad, como se ilustra en los ejemplos y explicaciones que siguen.

Para analizar más a fondo estos errores, extrajimos los pesos de atención asignados por cada uno de los cinco transformers a cada palabra de las instancias mal clasificadas. Estos pesos fueron sumados y usados para colorear los textos, destacando las palabras más influyentes según sus puntuaciones de atención.

A continuación, se presentan ejemplos ilustrativos de falsos positivos y falsos negativos:

Falso Positivo: <s>@usuario llevo mi cámara preparada para ser chica de artes q hace fotos a todos <s>

En este ejemplo, el modelo clasificó incorrectamente la declaración como sexista. La frase “ser chica” pudo haber sido marcada por el modelo como potencialmente indicativa de sexismo, especialmente cuando se considera junto con otros términos de la declaración. Sin embargo, aunque esta frase puede ser interpretada de diversas maneras, el contexto general sugiere que la declaración se refiere más a tomar fotos en un entorno social o artístico que a realizar un comentario sexista. La clasificación errónea probablemente ocurrió porque el modelo detectó un término potencialmente sensible pero no consideró adecuadamente el contexto, lo que llevó a un etiquetado incorrecto de sexismo.



Figura 12. Nube de Palabras Representando los Términos Comunes en los Tuits Clasificados como Falsos Positivos en el Modelo de Ensamble para Detectar Sexismo.

Generamos una nube de palabras, ilustrada en la Figura 12, para visualizar los términos más frecuentes en ejemplos de falsos positivos identificados por el modelo en el contexto de la detección de sexismo. Esta nube de palabras destaca los términos a los que el modelo prestó más atención en cada tuit, pero que erróneamente clasificó como sexistas. Términos prominentes como “mujer”, “sexual” y “preguntó” sugieren que las discusiones a menudo giraban en torno a temas de género y asuntos sexuales, áreas donde el modelo pudo haber malinterpretado el contexto o la intención. La aparición frecuente de términos relacionados con mujeres y sexualidad, como “mujer”, “sexual”, señala los desafíos para distinguir entre declaraciones neutrales o de apoyo sobre género y aquellas que realmente exhiben actitudes sexistas. Además, términos como “cabeza” y “ataque” pueden reflejar contextos en los que se utilizó un lenguaje fuerte o confrontativo, lo que llevó al modelo a clasificar incorrectamente el contenido como sexista. Sin embargo, es importante señalar que el significado real depende en gran medida del contexto en el que se usaron estos términos. A pesar de la presencia de estas palabras, su interpretación como sexista o no sexista puede variar significativamente dependiendo del contexto circundante, lo que probablemente contribuyó a la clasificación errónea en estos casos. Esta nube de palabras sirve como una herramienta

que aún clasificó incorrectamente como no sexistas. Términos prominentes como “violencia”, “chica” y “empresa” sugieren que las discusiones a menudo involucraban violencia e implicaciones laborales, áreas donde el modelo pudo haber tenido dificultades para detectar de manera precisa el sexismo sutil o dependiente del contexto. La frecuente aparición de términos relacionados con género y violencia, como “niña pequeña”, “chica” y “ayuda”, subraya los desafíos para detectar lenguaje potencialmente sexista, particularmente cuando se expresa de maneras indirectas o matizadas. Además, la presencia de términos como “color” puede indicar referencias a temas raciales o culturales que el modelo también encontró difíciles de clasificar correctamente en el contexto del sexismo.

Aunque el modelo ha mostrado un buen rendimiento general, la tendencia a generar falsos positivos con mayor frecuencia que falsos negativos resalta la necesidad de ajustar su sensibilidad y precisión, especialmente en contextos complejos y matizados. La incorporación de las puntuaciones de atención, desglosadas por subgrupos demográficos, resulta fundamental para comprender mejor las decisiones del modelo y afinar su desempeño, adaptándolo a diferentes grupos poblacionales. A medida que avanzamos, es crucial realizar ajustes adicionales que no solo optimicen la clasificación de sexismo, sino que también aumenten la transparencia del modelo y su capacidad para reconocer formas sutiles de discriminación. Esto permitirá un sistema más robusto, justo y eficaz en la detección de contenido sexista, con un mejor balance entre sensibilidad y especificidad, reduciendo errores y mejorando su rendimiento global.

En última instancia, este análisis detallado no solo contribuye a mejorar el rendimiento general del modelo, sino que también ayuda a construir un sistema más equitativo y transparente. Al comprender mejor las fuentes de error y las dinámicas de atención en las predicciones, podemos ajustar tanto los datos de entrenamiento como los algoritmos subyacentes para minimizar errores y asegurar una mayor precisión y equidad en la clasificación.

6.4. Explicaciones de los Tuits Clasificados con ChatGPT

El análisis y clasificación de tuits como sexistas o no sexistas constituye una tarea compleja, en la que la precisión del modelo debe ir acompañada de la transparencia en la toma de decisiones. Para lograr una mayor comprensión de cómo y por qué un modelo realiza ciertas clasificaciones, es fundamental contar con explicaciones claras y detalladas que permitan entender los criterios y las razones subyacentes a dichas decisiones. En este contexto, el uso de un modelo de lenguaje como GPT-4o resulta útil para proporcionar explicaciones claras y accesibles sobre las clasificaciones generadas. Aunque no es la única

alternativa posible, los modelos de lenguaje tienen la ventaja de permitir una mayor transparencia en la interpretación de los resultados. Esto facilita la identificación de los factores que influyen en la clasificación de los tuits. Sin embargo, es importante tener en cuenta que, aunque útiles, estos modelos pueden estar sujetos a errores y sesgos, lo que requiere una evaluación cuidadosa de sus respuestas.

Para ilustrar cómo funciona este proceso, se presentan a continuación dos ejemplos de tuits que han sido clasificados según su contenido: uno como sexista y el otro como no sexista. Estos ejemplos han sido seleccionados para mostrar las diferencias clave en los tipos de contenido que pueden llevar a cada clasificación, así como para proporcionar una explicación detallada del razonamiento detrás de la etiqueta asignada.

Tuit Sexista:

“@Diamalware @tuyomiespiritu La mujeres no pueden ser protectoras porque, por regla general, son más débiles físicamente que los hombres. Que algunas “mujeras” se traguen la propaganda globalista demuestra que su coeficiente intelectual está cerca de 80.”

Explicación:

Este tuit fue clasificado como sexista debido a que perpetúa varios estereotipos de género. En primer lugar, afirma que las mujeres son inherentemente más débiles físicamente que los hombres, lo que representa una generalización reductora basada en el género. Además, utiliza el término “mujeras”¹ (un término despectivo) y sugiere que las mujeres que aceptan ciertas ideas tienen un bajo coeficiente intelectual. Estas afirmaciones desvalorizan a las mujeres y refuerzan la idea de que son inferiores a los hombres tanto en capacidades físicas como intelectuales, lo cual es una forma clara de sexismo.

Tuit No Sexista:

“@HectorlCharles Si piensas que es de gente buena vs gente mala, qué mentalidad tan estrecha tienes.”

Explicación:

¹mujeras[sic]. El término correcto en español para referirse a mujeres es “mujeres”.

Este tuit no contiene elementos sexistas, ya que se limita a criticar la perspectiva o mentalidad de una persona sin hacer referencia al género ni perpetuar estereotipos de género. La crítica está dirigida a la simplificación del pensamiento de “gente buena vs gente mala” y no implica ningún juicio o comentario basado en el sexo o género de la persona a la que se dirige. Por lo tanto, se clasifica como no sexista.

Tuit Sexista pero clasificado No Sexista:

“@milf_alexxa Muy buenas tardes mi perra saludos desde Monterrey NL”

Explicación:

Aunque el tuit utiliza un término vulgar como “mi perra”, en este contexto parece más una expresión de confianza, camaradería o humor, que no necesariamente está ligada a perpetuar estereotipos de género o a un desvalor hacia las mujeres. Además, no hay un contexto que indique que esta frase sea denigrante o refuerce ideas sexistas explícitas. El modelo probablemente considera que este tuit no perpetúa desigualdades o discriminación basadas en el género.

Tuit No Sexista pero clasificado Sexista::

“@zaynsken Yo creo que de ser top 3 se habría ido contento pero lo de Jorginho le ha tocado la polla y con razón porque ni de coña ha hecho mejor año que él, lo de CR da puto cringe”

Explicación:

Este tuit incluye un lenguaje coloquial y vulgar que, aunque común en muchas conversaciones, presenta un tono problemático en el contexto del análisis de sexismo. La frase “le ha tocado la polla” puede perpetuar un lenguaje centrado en lo masculino de forma despectiva hacia el contexto competitivo en cuestión. Si bien no es un ejemplo evidente de sexismo directo, el uso de este tipo de lenguaje contribuye al refuerzo de un enfoque cultural que desvaloriza o trivializa elementos en favor de un enfoque masculinizado.

A lo largo del proceso de utilizar GPT-4o para generar las explicaciones, se presentaron varios desafíos que subrayan la importancia de estructurar un prompt claro y específico. En muchos casos, cuando el prompt no pedía una explicación detallada, el modelo simplemente devolvía la etiqueta asignada sin ofrecer información adicional sobre el proceso de clasificación. Esto nos llevó a incluir ejemplos explícitos dentro del prompt, lo que ayudó a guiar al modelo y asegurar que las explicaciones fueran completas tanto para

los tuits clasificados como sexistas como para los no sexistas. Sin embargo, notamos que el modelo tendía a omitir las explicaciones cuando los tuits eran clasificados como “no sexistas”, a menos que se solicitara específicamente. Además, las limitaciones de la versión gratuita de GPT-4o impidieron procesar todos los tuits a la vez, lo que nos obligó a dividir el conjunto de datos en partes más pequeñas. Finalmente, observamos que el modelo trataba las etiquetas proporcionadas como correctas sin cuestionarlas, lo que llevó a situaciones en las que se justificaban etiquetas incorrectas, como cuando un tuit sexista fue mal clasificado como “no sexista”.

El uso de GPT-4 para proporcionar explicaciones detalladas sobre las clasificaciones de los tuits ha mostrado ser una herramienta útil para comprender el proceso de toma de decisiones del modelo. Las explicaciones no solo ayudan a clarificar los criterios subyacentes que guían la clasificación de los tuits como sexistas o no sexistas, sino que también promueven la confianza en la efectividad y equidad del sistema. A pesar de los desafíos enfrentados, como la necesidad de un prompt bien estructurado y las limitaciones técnicas de la versión gratuita, el enfoque ha mostrado ser útil para mejorar tanto la precisión como la transparencia en la moderación automatizada de contenido. Este enfoque puede servir como base para futuras aplicaciones en modelos de clasificación de contenido, donde la explicación de las decisiones del modelo se vuelve crucial para garantizar su adopción y uso responsable.

6.5. Resumen del capítulo

En este capítulo hemos explorado la identificación automática del sexismo en redes sociales, poniendo especial énfasis en cómo las percepciones de los anotadores, diferenciados por género y edad, influyen en el desempeño de los modelos de clasificación.

Los hallazgos destacan la importancia de incorporar perspectivas diversas, tanto de género como de edad, en el desarrollo de sistemas para la detección automática de contenido ofensivo. En particular, los modelos que integran un amplio espectro de opiniones en sus consensos presentan un mejor rendimiento general. Sin embargo, las variaciones en las percepciones subrayan la necesidad de seguir refinando estos sistemas para captar las sutilezas del lenguaje y garantizar un análisis más inclusivo y matizado.

En conclusión, este capítulo demuestra que la percepción del sexismo es heterogénea y depende del perfil demográfico de los anotadores. Estos resultados tienen implicaciones significativas para el diseño de tecnologías de moderación de contenido más justas y precisas, subrayando la importancia de integrar

múltiples puntos de vista en el entrenamiento y desarrollo de modelos automáticos. Esto no solo mejora la eficacia de los sistemas, sino que también promueve una aproximación más equitativa y representativa al análisis del contenido digital.

Un aspecto clave en este sentido es el uso de herramientas como ChatGPT para generar explicaciones sobre las decisiones tomadas por los modelos de clasificación. Mediante la incorporación de explicaciones claras y comprensibles, se puede ofrecer un mayor entendimiento de cómo y por qué un sistema ha identificado determinado contenido como sexista, lo que contribuye a una mayor transparencia y confianza en los resultados. Esta capacidad de explicar las decisiones de los modelos no solo facilita la revisión y el ajuste de los mismos, sino que también refuerza la objetividad y la equidad en el análisis, proporcionando un enfoque más inclusivo y representativo del contenido digital.

Capítulo 7. Conclusiones

Este capítulo presenta una síntesis detallada de los principales hallazgos obtenidos en esta investigación sobre la detección automática de sexismo en español, destacando las diferencias demográficas en la percepción de este fenómeno, los avances metodológicos propuestos y las limitaciones encontradas a lo largo del estudio. Asimismo, se resaltan las contribuciones clave al campo del Procesamiento de Lenguaje Natural (PLN) y se proponen líneas futuras para superar los desafíos detectados, maximizar el impacto de los resultados y promover nuevas investigaciones en este ámbito.

7.1. Discusión

El análisis realizado en esta tesis pone de manifiesto hallazgos significativos en relación con las diferencias demográficas en la percepción del sexismo, particularmente en función de variables como género y edad. Un resultado destacado es la discrepancia del 25% entre las anotaciones realizadas por hombres y mujeres, lo que evidencia la subjetividad inherente al concepto de sexismo. Este hallazgo refuerza la importancia de integrar múltiples perspectivas en la construcción de modelos, garantizando su capacidad para reflejar una variedad más amplia de interpretaciones sociales y culturales.

Desde el punto de vista metodológico, la implementación de un modelo combinado basado en técnicas de ensamble demostró ser eficaz para capturar sensibilidades diversas respecto al sexismo. Aunque se identificaron limitaciones en su desempeño para ciertos grupos, como personas mayores de 46 años, este enfoque permitió integrar de manera efectiva diferentes perspectivas, mejorando el desempeño general del sistema. Además, el uso de herramientas como ChatGPT enriqueció el análisis, ofreciendo explicaciones interpretables que resultaron esenciales para identificar errores, comprender patrones y evaluar el impacto de las decisiones tomadas por los modelos.

En cuanto a la Tarea 1, los perfiles femeninos tendieron a identificar con mayor frecuencia elementos sexistas en los textos, mientras que los perfiles masculinos mostraron mayor tolerancia o ambigüedad hacia el mismo contenido. El enfoque de ensamble, que combinó predicciones de modelos individuales mediante votación mayoritaria, mejoró significativamente la robustez y precisión de las predicciones. Además, la segmentación por grupos de edad (18-22, 23-45 y 46+) evidenció que los usuarios más jóvenes eran más sensibles a identificar comportamientos sexistas, mientras que los mayores de 46 años presentaban patrones más conservadores o neutros. Sin embargo, este último grupo mostró limitaciones

en la representatividad del modelo, lo que sugiere la necesidad de ampliar las muestras y ajustar los hiperparámetros en futuros trabajos para garantizar un desempeño óptimo.

Respecto a la Tarea 2, uno de los hallazgos más relevantes fue la dificultad en generar una etiqueta única para el tipo de intención de los tuits sexistas, ya que no siempre los anotadores coincidían en la misma intención. Sin embargo, al implementar el voto basado en probabilidades, esta decisión mejoró significativamente, logrando una mayor precisión y consenso entre las predicciones del modelo.

Por otro lado, en la Tarea 3, al igual que en la Tarea 2, se presentaron complicaciones para alcanzar un consenso entre las categorías seleccionadas por los anotadores, incluso cuando esta tarea permitía elegir más de una categoría por tuit. En ocasiones, no se coincidía en ninguna categoría entre los anotadores, lo que dificultaba la consolidación de las etiquetas finales. Para abordar este desafío, la votación dura con enfoque probabilístico resultó ser esencial para manejar estas complejidades, permitiendo seleccionar una categoría siempre que fuese mencionada al menos una vez, lo que aseguró una mayor consistencia en los resultados.

Adicionalmente, se llevaron a cabo experimentos complementarios para analizar el impacto de las discrepancias demográficas. En estos, el modelo se entrenó utilizando datos de categorías específicas, como hombres, mujeres o grupos etarios (18-22 años), y se evaluó con etiquetas generadas a partir de opiniones de otros grupos, como personas mayores de 46 años o mujeres. Estos experimentos, ilustrado en el Anexo B, permitiendo observar las divergencias en las percepciones según el grupo de entrenamiento y prueba. Los resultados resaltaron patrones de contradicción o alineación entre las sensibilidades de los diferentes grupos, confirmando la relevancia de considerar dinámicas interdemográficas en el diseño y validación de los modelos.

En conjunto, estos hallazgos subrayan la importancia de adoptar enfoques metodológicos que consideren tanto las diferencias demográficas como las complejidades inherentes a los problemas sociales analizados. El uso de técnicas de ensamble y enfoques probabilísticos no solo mejoró la robustez del sistema, sino que también permitió capturar de manera más precisa las diversas perspectivas y patrones asociados con el sexismo en el lenguaje.

7.2. Limitaciones

Aunque los avances logrados son significativos, esta investigación presenta varias limitaciones que deben considerarse en futuras exploraciones.

En primer lugar, el corpus utilizado, aunque robusto y cuidadosamente seleccionado, está centrado exclusivamente en el idioma español y en un conjunto de datos específico. Esto limita la generalización de los resultados a otras lenguas o variantes regionales del español, donde las expresiones o los regionalismos podrían no estar representados.

En segundo lugar, las diferencias demográficas consideradas en este estudio se enfocaron en género y edad, dejando sin abordar otras variables relevantes como el nivel educativo, el contexto cultural y el nivel socioeconómico. Estas dimensiones podrían proporcionar una comprensión más amplia del fenómeno y revelar sesgos adicionales no explorados en este trabajo.

Además, el modelo combinado enfrentó desafíos técnicos, como una tasa de falsos positivos del 38.8 %, lo que subraya la necesidad de seguir perfeccionando las técnicas de clasificación, especialmente en contextos donde la precisión es crucial. Por último, aunque la incorporación de ChatGPT añadió valor al análisis, su uso podría introducir sesgos derivados de su propio entrenamiento o generar respuestas plausibles pero incorrectas, lo que representa un desafío para su aplicación práctica.

7.3. Contribuciones

Esta investigación ha realizado varias contribuciones significativas al campo del PLN y a la detección de sexismo:

- **Análisis exhaustivo de sesgos demográficos:** Se identificaron discrepancias importantes en la percepción del sexismo según género y edad, proporcionando información clave para el desarrollo de modelos más inclusivos y representativos.
- **Implementación de un modelo de ensamble:** Este enfoque superó en desempeño a los clasificadores individuales, logrando integrar perspectivas múltiples en un sistema único y cohesionado.
- **Avances en explicabilidad:** La incorporación de herramientas como ChatGPT facilitó la interpretación de los resultados, contribuyendo a su posible aplicación práctica en contextos como la moderación automatizada de contenido.
- **Corpus EXIST para estudios de sexismo:** Se realizaron mejoras significativas en el preprocesamiento y la segmentación del corpus, adaptándolo específicamente a la detección de sesgos por

diferentes grupos demográficos en español. Estas adaptaciones garantizaron una mayor representatividad y utilidad del conjunto de datos para experimentos posteriores, marcando un avance en la aplicación del PLN en temas de igualdad.

7.4. Artículos derivados de esta investigación

Como parte de los resultados obtenidos en esta investigación, se desarrollaron tres publicaciones científicas que reflejan los avances y hallazgos realizados:

1. **An Analysis of the Impact of Gender and Age on Perceiving and Identifying Sexist Posts:** Presentado en la Mexican Conference on Pattern Recognition (MCPR 2024), celebrada en Xalapa, Veracruz, México, del 19 al 22 de junio de 2024. Este artículo explora cómo las características demográficas de los anotadores, específicamente género y edad, afectan la percepción y clasificación de mensajes sexistas en español. Utilizando un conjunto de datos enriquecido con etiquetas provenientes de seis perfiles sociodemográficos diversos, se identificaron diferencias significativas en la interpretación del sexismo, con una discrepancia notable del 25 % entre las anotaciones de hombres y mujeres. Además, se destaca el desempeño del modelo RoBERTuito, que alcanza un puntaje F1 de 0.856 al ser entrenado únicamente con las etiquetas de anotadores del grupo de mayor edad, subrayando la importancia de desarrollar modelos que incorporen estas sensibilidades para una detección más inclusiva y precisa (Jimenez-Martinez et al., 2024a).
2. **Enhancing Sexism Detection in Tweets with Annotator-Integrated Ensemble Methods and Multimodal Embeddings for Memes:** Fue publicado en los Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), celebrado en Grenoble, Francia, del 9 al 12 de septiembre de 2024. En esta publicación, se describe nuestra participación en el reto EXIST 2024, centrado en la detección y categorización de mensajes sexistas en redes sociales y memes. Para los tuits, se combinan modelos transformers preentrenados con técnicas de ensamble, integrando perfiles de anotadores (como género y edad) para mejorar la identificación de sexismo y la intención del autor. En el caso de los memes, se utilizan embeddings CLIP generados por un modelo Vision Transformer (ViT) junto con clasificadores avanzados como redes neuronales y máquinas de factorización. Este enfoque multimodal e integrador obtuvo resultados destacados, incluyendo el primer lugar en la tarea de categorización de memes en inglés y español, demostrando la eficacia de incorporar factores demográficos y estrategias multimodales para robustecer la detección del

sexismo (Jimenez-Martinez et al., 2024b).

3. **Enhancing the Detection of Sexist Messages Through a Multi-Profile Based Ensemble**

Approach: Fue aceptado para su publicación en *Computación y Sistemas Journal* (Factor de impacto 0.6 según *Journal Citation Reports, JCR*). Este artículo aborda la detección de lenguaje sexista en español, un tema menos explorado en comparación con el inglés, con un enfoque en México, donde las normas de género tradicionales están profundamente arraigadas. A través del corpus EXIST, anotado por perfiles demográficos diversos, se identificaron discrepancias significativas en la percepción del sexismo, especialmente entre hombres y mujeres. Se propone un modelo de ensamble que combina clasificadores específicos para género y edad basados en RoBERTuito, logrando un puntaje F1 de 0.854. Además, se incluye un análisis de errores mediante visualización de pesos de atención y se utiliza ChatGPT para generar explicaciones interpretativas, destacando palabras clave y capturando matices lingüísticos. Este enfoque resalta la relevancia de integrar perspectivas diversas para mejorar la detección de mensajes sexistas en redes sociales.

Estos trabajos no solo evidencian el alcance y la calidad de la investigación, sino que también representan contribuciones clave al estado del arte en el procesamiento de lenguaje natural aplicado al análisis de mensajes discriminatorios.

7.5. Trabajo a Futuro

Con base en las limitaciones identificadas, se proponen las siguientes líneas de investigación futura:

- **Extensión lingüística y cultural:** Evaluar el desempeño de los modelos en diferentes variantes del español y extenderlos a otros idiomas, considerando las particularidades lingüísticas y culturales.
- **Inclusión de nuevos factores demográficos:** Incorporar variables como nivel educativo, clase social y contexto cultural para analizar cómo afectan la percepción del sexismo y los resultados de los modelos.
- **Mejoras en las arquitecturas:** Explorar el uso de transformers más avanzados para reducir la tasa de falsos positivos y mejorar la precisión general de los modelos.
- **Validación en entornos reales:** Implementar los modelos en plataformas dinámicas, como redes sociales, para evaluar su adaptabilidad a datos en tiempo real.

- **Refinamiento en explicabilidad:** Mejorar las herramientas de interpretabilidad para garantizar la veracidad y utilidad de las explicaciones generadas por los modelos.

Esta tesis aborda un problema de gran relevancia social y técnica, proponiendo enfoques innovadores que combinan precisión, inclusividad e interpretabilidad en la detección de sexismo en español. Las contribuciones realizadas constituyen una base sólida para investigaciones futuras y aplicaciones prácticas que busquen ampliar el impacto de estos modelos en el ámbito social y tecnológico.

Literatura citada

- Akhtar, S., Basile, V., & Patti, V. (2021). *Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection*. *arXiv preprint arXiv:2106.15896*, 1–24. <https://arxiv.org/pdf/2106.15896>.
- Anderson, L. & Barnes, M. (2022). *Hate speech*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/hate-speech/>.
- Böck, J., Schütz, M., Liakhovets, D., Satriani, N. Q., Babic, A., Slijepcevic, D., Zeppelzauer, M., & Schindler, A. (2023). *AIT_FHSTP at EXIST 2023 Benchmark: Sexism Detection by Transfer Learning, Sentiment and Toxicity Embeddings and Hand-Crafted Features*. In *CLEF (Working Notes)*, 878–890. <https://ceur-ws.org/Vol-3497/paper-074.pdf>.
- Comisión Nacional para Prevenir y Erradicar la Violencia Contra las Mujeres (2016). *¿Qué es el lenguaje sexista y por qué es importante visibilizarlo?* Gobierno de México. <https://www.gob.mx/conavim/articulos/que-es-el-lenguaje-sexista-y-por-que-es-importante-visibilizarlo?idiom=es>.
- Comisión Nacional para Prevenir y Erradicar la Violencia Contra las Mujeres (2018). *Frases sexistas que hombres y mujeres debemos dejar de decir para promover la igualdad de género*. Gobierno de México. <https://www.gob.mx/conavim/articulos/frases-sexistas-que-hombres-y-mujeres-debemos-dejar-de-decir-para-promover-la-igualdad-de-genero>.
- de la Morena, Ines (2020). *Machismo, Femicides, and Child's Play: Gender Violence in Mexico*. Harvard International Review. <https://hir.harvard.edu/gender-violence-in-mexico-machismo-femicides-and-childs-play/>.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. *arXiv preprint arXiv:1810.04805*. <https://arxiv.org/pdf/1810.04805>.
- Diego Olite, F. M., Morales Suárez, I. d. R., & Vidal Ledo, M. J. (2023). *Chat GPT: origen, evolución, retos e impactos en la educación*. *Educación Médica Superior*, 37(2). http://scielo.sld.cu/scielo.php?pid=S0864-21412023000200016&script=sci_arttext.
- Fang, Y.-Z., Lee, L.-H., & Huang, J.-D. (2024). *NYCU-NLP at EXIST 2024—leveraging transformers with diverse annotations for sexism identification in social networks*. *Working Notes of CLEF*. <https://ceur-ws.org/Vol-3740/paper-93.pdf>.
- Ferrer, J. (2024). *How Transformers Work: A Detailed Exploration of Transformer Architecture*. DataCamp. <https://www.datacamp.com/tutorial/how-transformers-work>.
- Fulper, R., Ciampaglia, G. L., Ferrara, E., Ahn, Y., Flammini, A., Menczer, F., Lewis, B., & Rowe, K. (2014). *Misogynistic language on Twitter and sexual violence*. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*, 57–64. Association for Computing Machinery. <https://yongyeol.com/papers/fulper2014misogynistic.pdf>.
- GeeksforGeeks (2024). *F1 Score in Machine Learning*. Geeksforgeeks. <https://www.geeksforgeeks.org/f1-score-in-machine-learning/>.
- Jahan, M. S. & Oussalah, M. (2023). *A systematic review of hate speech automatic detection using natural language processing*. *Neurocomputing*, 546, 126232. <https://www.sciencedirect.com/science/article/pii/S0925231223003557>.
- Jain, S. & Wallace, B. C. (2019). *Attention is not explanation*. *arXiv preprint arXiv:1902.10186*. <https://arxiv.org/pdf/1902.10186>.

- Jimenez-Martinez, M. P., Lopez-Nava, I. H., & Montes-y Gómez, M. (2024a). *An Analysis of the Impact of Gender and Age on Perceiving and Identifying Sexist Posts*. In *Mexican Conference on Pattern Recognition*, 308–318. Springer. https://link.springer.com/chapter/10.1007/978-3-031-62836-8_29.
- Jimenez-Martinez, M. P., Raygoza-Romero, J. M., Sánchez-Torres, C. E., Lopez-Nava, I. H., & Montes-y Gómez, M. (2024b). *Enhancing sexism detection in tweets with annotator-integrated ensemble methods and multimodal embeddings for memes*. *Working Notes of CLEF*. <https://ceur-ws.org/Vol-3740/paper-97.pdf>.
- Koonireddy, R. & Adel, N. (2023). *ROH_NEIL@ EXIST2023: Detecting Sexism in Tweets using Multilingual Language Models*. In *CLEF (Working Notes)*, 967–984. <https://ceur-ws.org/Vol-3497/paper-083.pdf>.
- Muti, A., Mancini, E., et al. (2023). *Enriching hate-tuned transformer-based embeddings with emotions for the categorization of sexism*. In *CEUR Workshop Proceedings*, volume 3497, 1012–1023. CEUR-WS. <https://cris.unibo.it/bitstream/11585/949866/1/paper-086.pdf>.
- Ocampo, N. B., Sviridova, E., Cabrio, E., & Villata, S. (2023). *An in-depth analysis of implicit and subtle hate speech messages*. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1997–2013. Association for Computational Linguistics. <https://hal.science/hal-04214094/document>.
- OpenAI (2022). *Introducing ChatGPT*. OpenAI. <https://openai.com/index/chatgpt/>.
- Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). *A survey of text representation and embedding techniques in nlp*. *IEEE Access*, 11, 36120–36146. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10098736>.
- Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). *Hate speech: A systematized review*. *Sage Open*, 10(4), 2158244020973022. <https://journals.sagepub.com/doi/pdf/10.1177/2158244020973022>.
- Pérez, J. M., Furman, D. A., Alemany, L. A., & Luque, F. (2021). *RoBERTuito: a pre-trained language model for social media text in Spanish*. *arXiv preprint arXiv:2111.09453*. <https://arxiv.org/pdf/2111.09453>.
- Plaza, L., Carrillo-de Albornoz, J., Morante, R., Amigó, E., Gonzalo, J., Spina, D., & Rosso, P. (2023). *Overview of exist 2023—learning with disagreement for sexism identification and characterization*. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 316–342. Springer. <https://ceur-ws.org/Vol-3497/paper-070.pdf>.
- Plaza, L., Carrillo-de Albornoz, J., Ruiz, V., Maeso, A., Chulvi, B., Rosso, P., Amigó, E., Gonzalo, J., Morante, R., & Spina, D. (2024). *Overview of EXIST 2024—Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes*. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 93–117. Springer. <https://www.damianospina.com/publication/plaza-2024-overview/plaza-2024-overview.pdf>.
- Plaza Del Arco, F. M. (2023). *Detecting offensive language by integrating multiple linguistic phenomena*. *Jaén: Universidad de Jaén*. https://ruja.ujaen.es/jspui/bitstream/10953/2400/1/Thesis_deposito%20FLOR%20MIRIAM.pdf.
- Pombo Nartallo, V. (2024). *¿Qué es la explicabilidad de la inteligencia artificial? Cómo quitarle misterio a la tecnología*. BBVA. <https://www.bbva.com/es/innovacion/que-es-la-explicabilidad-de-la-ia-como-quitarle-misterio-a-la-tecnologia/>.

- Quan, L. M. & Thin, D. V. (2024). *Sexism Identification in Social Networks with Generation-based Language Models*. Working Notes of CLEF. <https://ceur-ws.org/Vol-3740/paper-109.pdf>.
- Real Academia Española (2024). *Sexismo*. <https://dle.rae.es/sexismo>.
- Reddy, S. (2023). *Transformer models and BERT model: Descripción general*. Google Cloud. <https://www.coursera.org/learn/transformer-models-and-bert-model---espaol/lecture/mM5NM/modelos-de-transformadores-y-modelo-bert-descripcion-general>.
- Richards, D., Caldwell, P. H., & Go, H. (2015). *Impact of social media on the health of children and young people*. *Journal of paediatrics and child health*, 51(12), 1152–1157. <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpc.13023>.
- Roselli, D., Matthews, J., & Talagala, N. (2019). *Managing bias in AI*. In *Companion proceedings of the 2019 world wide web conference*, 539–544. https://lin-web.clarkson.edu/~jmatthew/publications/ManagingBiasInAI_CAMERAREADY.pdf.
- Secretaría de Salud (2011). *Manual para el uso no sexista del lenguaje*. CNEGSR. Gobierno de México. <https://www.gob.mx/salud/documentos/manual-para-el-uso-no-sexista-del-lenguaje-cnegsr>.
- Singh, S. K., Kumar, S., & Mehra, P. S. (2023). *Chat gpt & google bard ai: A review*. In *2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*, 1–6. IEEE. https://www.researchgate.net/profile/Pawan-Mehra/publication/374398713_Chat_GPT_Google_Bard_AI_A_Review/links/651ea65cb0df2f20a21392da/Chat-GPT-Google-Bard-AI-A-Review.pdf.
- Tavarez-Rodríguez, J., Sánchez-Vega, F., Rosales-Pérez, A., & López-Monroy, A. P. (2024). *Better together: LLM and neural classification transformers to detect sexism*. Working Notes of CLEF. <https://ceur-ws.org/Vol-3740/paper-118.pdf>.
- Tian, L., Huang, N., & Zhang, X. (2023). *Efficient Multilingual Sexism Detection via Large Language Model Cascades*. In *CLEF (Working Notes)*, 1083–1090. <https://ceur-ws.org/Vol-3497/paper-092.pdf>.
- Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers*. O'Reilly Media, Inc. <https://www.oreilly.com/library/view/natural-language-processing/9781098136789/>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. *Advances in neural information processing systems*, 30. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- Venegas-Vera, A. V., Colbert, G. B., & Lerma, E. V. (2020). *Positive and negative impact of social media in the COVID-19 era*. *Reviews in cardiovascular medicine*, 21(4), 561–564. <https://doi.org/10.31083/j.rcm.2020.04.195>.
- Villarreal-Haro, K., Sánchez-Vega, F., Rosales-Pérez, A., & López-Monroy, A. P. (2024). *Stacked reflective reasoning in large neural language models*. Working Notes of CLEF. <https://ceur-ws.org/Vol-3740/paper-121.pdf>.
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). *Hierarchical attention networks for document classification*. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489. <https://aclanthology.org/N16-1174.pdf>.

Anexos

Anexo A. Puntaje F1 macro

El puntaje F1 macro es una métrica en aprendizaje automático que combina precisión y exhaustividad (recall) para evaluar el rendimiento del modelo, especialmente en problemas multiclase.

- **Precisión:** Mide qué proporción de las predicciones positivas del modelo son realmente correctas:

$$\text{Precisión} = \frac{TP}{TP + FP} \quad (1)$$

donde TP = Verdaderos Positivos y FP = Falsos Positivos.

- **Recall (Exhaustividad):** Mide qué proporción de los casos positivos reales fueron identificados correctamente:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

donde FN = Falsos Negativos.

- **F1 Score:** Combina precisión y recall en una única métrica balanceada:

$$F1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}} \quad (3)$$

El puntaje F1 macro calcula el F1 Score para cada clase por separado y luego obtiene el promedio aritmético:

$$F1_{macro} = \frac{1}{N} \sum_{i=1}^N F1_i \quad (4)$$

- N = Número total de clases.
- $F1_i$ = F1 Score de la clase i .

Este método no toma en cuenta el desequilibrio de clases, ya que trata a todas las clases por igual, independientemente de su tamaño (GeeksforGeeks, 2024).

Anexo B. Análisis de desempeño del modelo según grupos demográficos y congruencia entre entrenamiento y prueba

También se realizaron experimentos adicionales utilizando nuestro modelo con distintos tipos de combinaciones de entrenamiento y prueba. Específicamente, el modelo se entrenó utilizando categorías demográficas específicas, como mujeres, hombres, y diferentes rangos de edad (por ejemplo, 18-22 años). Posteriormente, se evaluó utilizando etiquetas generadas a partir de las opiniones de otros grupos demográficos, como personas mayores de 46 años o mujeres. Este enfoque permitió analizar el grado de discrepancia o contradicción entre las percepciones de los diferentes grupos demográficos.

De acuerdo con la Figura 14, los resultados del análisis revelan tendencias interesantes en el desempeño de los diferentes grupos de edad. Al evaluar los grupos utilizando las opiniones del rango de edad de 18 a 22 años, el mejor F1 score, con un valor de 0.8800, se obtuvo en el grupo de 46 años en adelante. Por otro lado, cuando se probaron los modelos con las opiniones del rango de edad de 23 a 45 años, el grupo que presentó el mejor desempeño fue el de 18 a 22 años, alcanzando un F1 score de 0.8653. Finalmente, al utilizar como prueba las opiniones del grupo de 46 años en adelante, el mayor F1 score fue nuevamente del grupo de 46 años en adelante, con un valor de 0.8753. Estos resultados sugieren que el grupo de 46 años en adelante muestra una mayor congruencia entre su entrenamiento y sus predicciones, destacándose tanto al ser evaluado con sus propias opiniones como con las de otros grupos.

Según la Figura 15, el grupo con mejor desempeño fue consistentemente el correspondiente a la etiqueta del test, destacando en todas las pruebas realizadas. Al ser evaluado con la etiqueta dura, así como con las categorías de hombres de 18 a 22 años, de 23 a 45 años, y de 46 años en adelante, este grupo alcanzó los F1 scores más altos, con valores de 0.847, 0.817, 0.793 y 0.799, respectivamente. Por otro lado, el segundo mejor desempeño lo presentó el grupo de mujeres de 23 a 45 años, que también mantuvo una posición destacada en todas las perspectivas del test. Este grupo obtuvo un F1 score de 0.798 con la etiqueta dura, 0.782 al ser probado con hombres de 18 a 22 años, 0.735 con hombres de 23 a 45 años y 0.733 con hombres de 46 años en adelante.

El grupo correspondiente a la etiqueta del test tiene una robustez significativa, siendo el más consistente en cada evaluación realizada. Además, aunque el grupo de mujeres de 23 a 45 años no alcanza el mejor desempeño, mantiene una notable estabilidad, lo que sugiere que las características propias de este grupo también lo posicionan como un competidor cercano en términos de desempeño. La diferencia entre los mejores F1 scores indica que los modelos tienden a adaptarse mejor al grupo objetivo del test, mostrando que la concordancia entre el entrenamiento y la prueba desempeña un papel crucial en los resultados

obtenidos.

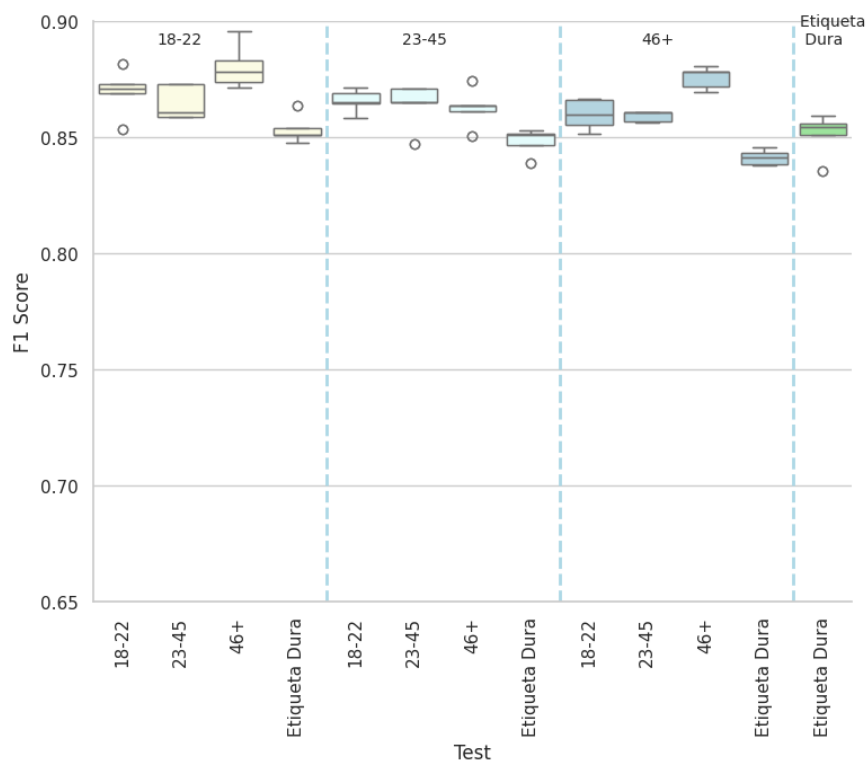


Figura 14. Cruces por edad.

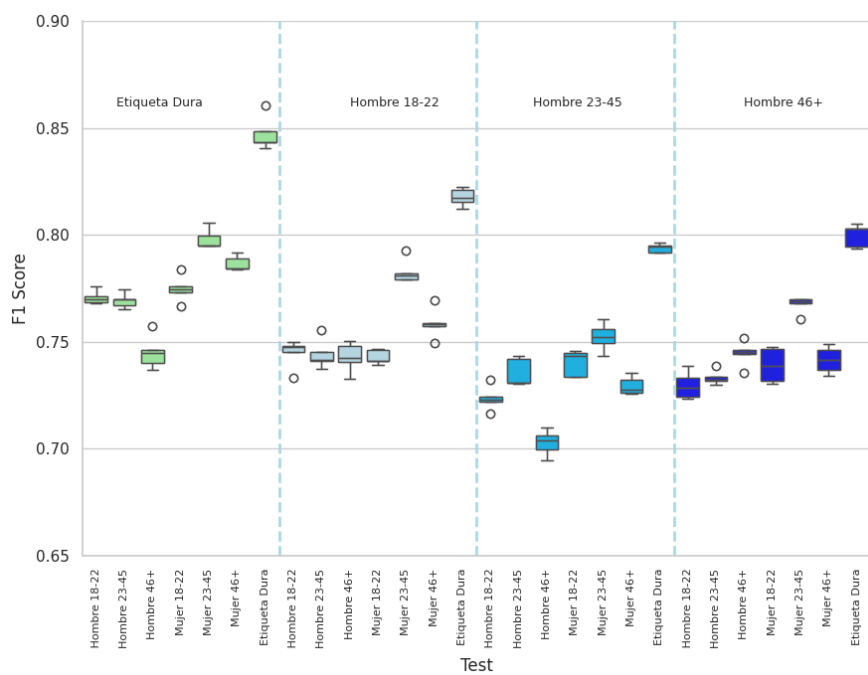


Figura 15. Cruces de hombres por distintas edades.

De la Figura 16, al igual que en el análisis realizado con el test de hombres, se evaluó el desempeño

del perfil de anotador frente a las opiniones de mujeres agrupadas por rango de edad. En este caso, el mejor desempeño en términos de F1 score correspondió a la etiqueta dura, destacando en las categorías de mujeres de 18 a 22 años, de 23 a 45 años y de 46 años en adelante, con F1 scores de 0.793, 0.816 y 0.800, respectivamente.

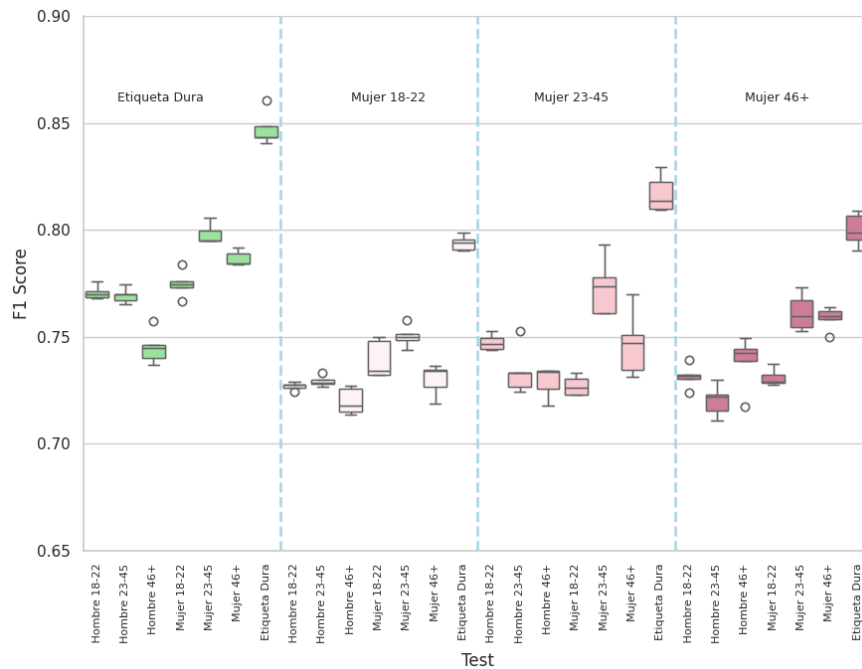


Figura 16. Cruces de mujeres por distintas edades.

El segundo mejor desempeño, para cada test, fue consistentemente del grupo de mujeres de 23 a 45 años. Este grupo obtuvo un F1 score de 0.750 al ser probado con mujeres de 18 a 22 años, 0.734 con mujeres de su misma categoría (23 a 45 años), y 0.761 al ser evaluado con mujeres de 46 años en adelante.

Estos resultados indican que la etiqueta dura sigue siendo la referencia con mejor desempeño, mostrando una notable consistencia al liderar en todas las categorías evaluadas. Por su parte, el grupo de mujeres de 23 a 45 años demuestra ser un perfil sólido y versátil, posicionándose como el segundo mejor grupo en cada evaluación. Este comportamiento sugiere que, al igual que en el test de hombres, la congruencia entre el perfil de entrenamiento y el rango de evaluación influye significativamente en los resultados, y que ciertos grupos de edad presentan características que los hacen más robustos y adaptables frente a diferentes perspectivas.

Los análisis realizados en este anexo destacan la importancia de la congruencia entre los perfiles de entrenamiento y las etiquetas de prueba, evidenciando patrones de desempeño específicos según los grupos demográficos evaluados. En general, los modelos muestran un mejor desempeño cuando el grupo de

prueba coincide con el grupo de entrenamiento, lo que sugiere una fuerte influencia de las características demográficas en la capacidad de predicción del modelo. Por otro lado, el perfil de mujer de 23 a 45 años y el perfil de personas mayores de 46 años destacan como perfiles consistentes y versátiles, posicionándose como los de mayor robustez frente a distintas pruebas. Estos resultados subrayan la necesidad de considerar la diversidad demográfica en el diseño y la evaluación de modelos, con el fin de garantizar su equidad y adaptabilidad en aplicaciones prácticas.