



UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

**DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA
INFORMACIÓN**

**REDES NEURONALES UNIMODALES PARA CLASIFICAR
MENSAJES MISÓGINOS EN REDES SOCIALES**

**TESIS PARA OBTENER EL GRADO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN**

PRESENTA:

ITZELY VIRIDIANA MEDINA TORRES

BAJO LA DIRECCIÓN DE:

DR. PABLO PANCARDO GARCÍA

CUNDUACÁN, TABASCO, A: JULIO 2026



UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA
INFORMACIÓN

**REDES NEURONALES UNIMODALES PARA CLASIFICAR
MENSAJES MISÓGINOS EN REDES SOCIALES**

TESIS PARA OBTENER EL GRADO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

ITZELY VIRIDIANA MEDINA TORRES

BAJO LA DIRECCIÓN DE:

DR. PABLO PANCARDO GARCÍA

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán el día tres del mes de Julio del año 2026, el que suscribe **Itzely Viridiana Medina Torres**, alumno del Programa de la **Maestría en Ciencias de la Computación** con número de matrícula **242H21001**, adscrito a la **División Académica de Ciencias y Tecnologías de la Información**, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención de Grado de Maestría y titulada **Redes neuronales unimodales para clasificar mensajes misóginos en redes sociales**, dirigida por el Dr. Pablo Pancardo García.

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Cunduacán, Tabasco a 03 de Julio de 2026.



Estudiante: Itzely Viridiana Medina Torres



UJAT

UNIVERSIDAD JUÁREZ
AUTÓNOMA DE TABASCO

"ESTUDIO EN LA DUDA, ACCIÓN EN LA FE"



DIVISIÓN ACADÉMICA DE
CIENCIAS Y TECNOLOGÍAS
DE LA INFORMACIÓN



2026
año de
Margarita
Maza

Cunduacán, Tabasco, a 03 de julio de 2026
Oficio No. 1404/2026/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Itzely Viridiana Medina Torres

Egresada de la Maestría en Ciencias de la Computación

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional "**Redes neuronales unimodales para clasificar mensajes misóginos en redes sociales**", para presentar examen y obtener el Grado de Maestra en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

UNIVERSIDAD JUÁREZ
AUTÓNOMA DE TABASCO

Atentamente

Dra. Laura Beatriz Vidal Turrubiates
Directora



DIVISIÓN ACADÉMICA DE
CIENCIAS Y TECNOLOGÍAS
DE LA INFORMACIÓN

C.c. p. Dr. Eduardo Hernández de la Cruz. - Encargado del despacho de la Coordinación de Posgrado.
Archivo

DRA.*LBVT/EHC

Av. Universidad s/n, Zona de la Cultura, Col. Magisterial,
Villahermosa, Centro, Tabasco, Mex. C.P. 86040.
Tel (993) 358 15 00 e-Mail: rectoria@ujat.mx

Carta de Cesión de Derechos

Villahermosa, Tabasco a 03 de Julio de 2026.

Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Redes neuronales unimodales para clasificar mensajes misóginos en redes sociales.**

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención de que tenemos el derecho al reconocimiento como autores de la obra, y a la Universidad Juárez Autónoma de Tabasco mantendrá en un 100 % la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR



Estudiante: Itzely Viridiana Medina Torres

TESTIGOS



Dr. Pablo Pancardo García



Dr. José Adán Hernández Nolasco

Agradecimientos

A Dios, por concederme la fortaleza, la sabiduría, la perseverancia y salud para superar cada desafío y hacer posible la culminación de esta etapa académica.

A la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), por el apoyo brindado mediante la beca de maestría, la cual hizo posible dedicar el tiempo y el esfuerzo necesarios para el desarrollo de esta investigación.

A mi director de tesis, por su orientación, conocimiento y compromiso durante el desarrollo de esta investigación. Sus observaciones y acompañamiento fueron fundamentales para fortalecer este trabajo.

A mis docentes, por compartir generosamente sus conocimientos, fomentar el pensamiento crítico e inspirarme a seguir creciendo como profesionista e investigadora.

A mi madre, por su apoyo constante e incondicional. Gracias por estar presente en cada etapa de este camino. Su ayuda fue fundamental para que pudiera dedicar el tiempo y el esfuerzo que esta meta demandó.

A mis hijas, quienes son siempre mi mayor motivación y la razón que impulsa cada uno de mis esfuerzos. Su amor, paciencia y comprensión dieron sentido a cada momento de este camino.

Finalmente, expreso mi más sincero agradecimiento a todas las personas que, de una u otra manera, contribuyeron a la realización de esta investigación. Este logro también refleja el apoyo, la confianza y las enseñanzas que recibí a lo largo de este proceso.

Índice general

Índice de Figuras	V
Índice de Tablas	VII
Resumen	VIII
Abstract	IX
1. Generalidades	1
1.1. Introducción	1
1.2. Planteamiento del problema	4
1.2.1. Definición del problema	4
1.2.2. Delimitación de la investigación	6
1.3. Pregunta de investigación e hipótesis	6
1.4. Objetivo general	6
1.5. Objetivos específicos	7
1.6. Justificación	7
1.7. Metodología utilizada	8
2. Marco teórico	11
2.1. Introducción	11
2.2. Conceptos y teorías fundamentales de la investigación	12
2.2.1. Redes neuronales artificiales	12
2.2.2. Redes neuronales unimodales versus multimodales	13

2.2.3.	Arquitecturas de redes neuronales para procesamiento de texto	14
2.2.3.1.	Redes Neuronales Recurrentes (RNN) y LSTM	14
2.2.3.2.	Redes Neuronales Convolucionales para texto	15
2.2.3.3.	Modelos basados en Transformers	16
2.2.4.	Procesamiento del Lenguaje Natural	17
2.2.4.1.	Preprocesamiento de texto	17
2.2.4.2.	Representaciones de palabras: Embeddings	18
2.2.5.	Métricas de evaluación en clasificación de texto	18
2.2.6.	Modelos de clasificación de discurso de odio	19
2.3.	Literatura relacionada	20
2.3.1.	Trabajos fundacionales en detección de discurso de odio	20
2.3.2.	Investigaciones en detección de misoginia	21
2.3.3.	Detección de discurso de odio en español	22
2.3.4.	Aplicaciones de redes neuronales en clasificación de texto en español	22
2.3.5.	Modelos de lenguaje preentrenados para español	23
2.3.6.	Desafíos y consideraciones éticas	24
2.3.7.	Perspectiva sociológica y cultural	24
2.3.8.	Análisis comparativo del estado del arte	25
2.4.	Marco tecnológico	26
2.4.1.	Software y frameworks	26
2.4.2.	Recursos lingüísticos y datasets	27
2.4.3.	Infraestructura computacional	28
2.4.4.	Hardware local	28
2.4.5.	Herramientas de control de versiones y documentación	28
2.4.6.	Justificación de la selección tecnológica	29
3.	Análisis exploratorio de datos	30
3.1.	Introducción	30
3.2.	Recopilación de datos	30
3.2.1.	Dataset MEX-A3T	30

3.2.2. Dataset EXIST 2024	32
3.2.3. Integración de los datasets	33
3.3. Preprocesamiento de texto	34
3.3.1. Eliminación de registros duplicados	34
3.3.2. Eliminación de registros incompletos	36
3.3.3. Conversión a minúsculas	37
3.3.4. Eliminación de URLs	38
3.3.5. Eliminación de menciones	40
3.3.6. Eliminación de caracteres especiales	41
3.4. Análisis estadístico	45
3.4.1. Distribución de clases	45
4. Experimentos y Resultados	48
4.1. Diseño experimental	48
4.2. Configuración de modelos	49
4.2.1. Modelos secuenciales	49
4.2.2. Modelos basados en Transformers	50
4.3. Entorno de ejecución	50
4.4. Métricas de evaluación	51
4.5. Resultados experimentales	52
4.5.1. Observación sobre sobreajuste en experimentos exploratorios	52
4.5.2. Fase 1: Resultados con datasets individuales	52
4.5.2.1. Resultados en el dataset MEX-A3T	52
4.5.2.2. Resultados en el dataset EXIST 2024	54
4.5.2.3. Análisis comparativo de la Fase 1	55
4.5.3. Fase 2: Resultados con dataset fusionado	55
4.5.3.1. Configuración experimental del dataset fusionado	55
4.5.3.2. Modelo BETO	56
4.5.3.3. Modelo XLM-RoBERTa large	57
4.5.3.4. Modelo CNN	58

4.5.3.5. Modelo LSTM	59
4.5.3.6. Análisis comparativo de la Fase 2	59
4.5.4. Comparación detallada: BETO, XLM-RoBERTa large y CNN	60
4.5.5. Impacto del dataset en el rendimiento	60
4.5.6. Especialización lingüística vs. capacidad paramétrica	61
4.5.7. Jerarquía de arquitecturas	61
5. Discusión, conclusiones y recomendaciones	63
5.1. Discusión	63
5.1.1. Limitaciones	65
5.2. Conclusiones	66
5.2.1. Conclusión final	67
5.3. Recomendaciones	67
5.3.1. Mejoras a corto plazo	67
5.3.2. Extensiones a mediano plazo	68
5.3.3. Investigación a largo plazo	68
Anexo	70
Bibliografía	71

Índice de figuras

3.1. Análisis de registros duplicados en el dataset MEX-A3T	35
3.2. Análisis de registros duplicados en el dataset EXIST 2024.	35
3.3. Verificación de registros incompletos en el dataset MEX-A3T.	37
3.4. Verificación de registros incompletos en el dataset EXIST 2024.	37
3.5. Estadísticas de conversión a minúsculas en el dataset MEX-A3T.	38
3.6. Estadísticas de conversión a minúsculas en el dataset EXIST 2024.	38
3.7. Análisis y eliminación de URLs en el dataset MEX-A3T.	40
3.8. Análisis y eliminación de URLs en el dataset EXIST 2024.	41
3.9. Análisis y eliminación de menciones en el dataset MEX-A3T.	42
3.10. Análisis y eliminación de menciones en el dataset EXIST 2024.	43
3.11. Análisis y eliminación de caracteres especiales en el dataset MEX-A3T.	44
3.12. Análisis y eliminación de caracteres especiales en el dataset EXIST 2024.	44
3.13. Distribución de clases en el dataset MEX-A3T. Se observa una distribución relativamente balanceada con 52.3 % de mensajes no misóginos y 47.7 % de mensajes misóginos.	46
3.14. Distribución de clases en el dataset EXIST 2024. La distribución muestra 59.7 % de contenido no sexista y 40.3 % de contenido sexista.	47

- 4.1. Curvas de entrenamiento del modelo BETO con el dataset fusionado. El modelo alcanza convergencia estable en 3 épocas con descenso consistente de la pérdida de entrenamiento hasta 0.15 y de validación hasta 0.24, sin divergencia entre ambas curvas. La accuracy de validación se mantiene estable en 88 % desde la primera época. Las métricas de validación muestran que el Macro-F1 supera claramente el umbral del 85 % en la época final, con Precision 95 % y Recall 89 %, confirmando la verificación de la hipótesis planteada. La línea punteada indica el umbral de 85 % establecido en la hipótesis. 57
- 4.2. Curvas de entrenamiento del modelo XLM-RoBERTa large con el dataset fusionado. El modelo converge en tan solo 3 épocas con una reducción limpia y estable de la pérdida tanto en entrenamiento como en validación, sin señales de sobreajuste. La accuracy de validación se estabiliza en 87 % y las métricas de validación muestran una Precision consistentemente alta (95 %) a lo largo de todo el entrenamiento. El Macro-F1 muestra una tendencia creciente que se aproxima al umbral del 85 % en la época final, sin llegar a superarlo de forma consolidada, lo que es consistente con el resultado reportado en la Tabla 4.9 (Macro-F1 = 84.13 %). . . . 58
- 4.3. Curvas de entrenamiento del modelo CNN con el dataset fusionado. La gráfica de pérdida muestra una reducción consistente del error de entrenamiento a lo largo de las 35 épocas, mientras que la pérdida de validación desciende con mayor variabilidad, evidenciando un sobreajuste leve pero controlado por los mecanismos de *dropout* y *early stopping*. La accuracy de validación se estabiliza en torno al 82 %, mientras que las métricas de validación muestran una Precision alta y estable (93 %) con un Recall que mejora progresivamente hasta aproximadamente 82 %, reflejando el efecto de los pesos de clase proporcionales sobre la detección de la clase minoritaria. 59

Índice de tablas

2.1. Comparación de trabajos previos en detección de misoginia y discurso de odio . . .	25
4.1. Resultados del modelo BETO entrenado con el dataset MEX-A3T.	53
4.2. Resultados del modelo CNN entrenado con el dataset MEX-A3T.	53
4.3. Resultados del modelo LSTM entrenado con el dataset MEX-A3T.	53
4.4. Resultados del modelo BETO entrenado con el dataset EXIST 2024.	54
4.5. Resultados del modelo CNN entrenado con el dataset EXIST 2024.	54
4.6. Resultados del modelo LSTM entrenado con el dataset EXIST 2024.	55
4.7. Comparación de Macro-F1 en la Fase 1 (datasets individuales).	55
4.8. Resultados del modelo BETO entrenado con el dataset fusionado.	56
4.9. Resultados del modelo XLM-RoBERTa large entrenado con el dataset fusionado. . .	57
4.10. Resultados del modelo CNN entrenado con el dataset fusionado.	58
4.11. Comparación de resultados entre arquitecturas con dataset fusionado.	59
4.12. Comparación detallada entre los modelos con dataset fusionado.	60

Resumen

Esta investigación desarrolla modelos basados en redes neuronales unimodales para clasificar mensajes misóginos en redes sociales, específicamente en español mexicano. Se evaluaron tres arquitecturas: LSTM, CNN y BETO (modelo transformer preentrenado en español) y se incorporó también XLM-RoBERTa large. Los modelos se entrenaron con dos conjuntos de datos: MEX-A3T (3,307 tweets) y EXIST 2024 (3,660 tweets), posteriormente fusionados para aprovechar su diversidad. La metodología siguió seis fases: Selección del modelo neuronal para textos en español, recolección y preprocesamiento de datasets, entrenamiento de los modelos, evaluación del rendimiento, implementación y validación para asegurar la generalización a datos no vistos. Los resultados demuestran que BETO superó el 85 % de exactitud en la detección de contenido misógino, confirmando la hipótesis inicial. BETO alcanzó 88.42 % de accuracy, y 95.13 % de precisión, superando la meta establecida. XLM-RoBERTa large obtuvo 87.45 % de accuracy. CNN se ubicó por debajo del objetivo con 81.85 %, LSTM no logró convergencia en el corpus fusionado, aunque alcanzó 74.19 % y 72.31 % de Macro-F1 en los datasets individuales. La fusión de datasets mejoró el rendimiento de BETO en 3.19 % puntos porcentuales. Los hallazgos confirman que las redes neuronales especializadas en español son efectivas para detectar patrones lingüísticos asociados a la misoginia. BETO demostró menor cantidad de falsos positivos (29 vs 41 de CNN), haciéndolo más confiable cuando la precisión es crítica. CNN, ofrece una alternativa eficiente con 78 veces menos parámetros y menor tiempo de entrenamiento. Este trabajo aporta evidencia empírica sobre la efectividad de modelos unimodales para clasificar discurso de odio en español mexicano, y ofrece herramientas prácticas para sistemas de moderación de contenido para crear espacios digitales más seguros e inclusivos.

Palabras Clave: Redes Neuronales Unimodales, redes sociales, misoginia.

Abstract

This research develops unimodal neural network-based models to classify misogynistic messages on social media, specifically in Mexican Spanish. Three architectures were evaluated: LSTM, CNN, and BETO (a transformer model pretrained in Spanish), and XLM-RoBERTa large was also incorporated. The models were trained using two datasets: MEX-A3T (3,307 tweets) and EXIST 2024 (3,660 tweets), which were subsequently merged to leverage their diversity.

The methodology followed six phases: selection of the neural model for Spanish texts, dataset collection and preprocessing, model training, performance evaluation, implementation, and validation to ensure generalization to unseen data.

The results demonstrate that BETO exceeded 85% accuracy in detecting misogynistic content, confirming the initial hypothesis. BETO achieved 88.42% accuracy and 95.13% precision, surpassing the established target. XLM-RoBERTa large achieved 87.45% accuracy. CNN fell below the target at 81.85%, and LSTM did not achieve convergence on the merged corpus, although it reached 74.19% and 72.31% Macro-F1 on the individual datasets. Merging the datasets improved BETO's performance by 3.19 percentage points.

The findings confirm that neural networks specialized in Spanish are effective at identifying linguistic patterns associated with misogyny. BETO produced fewer false positives (29 vs. 41 for CNN), making it more reliable when high precision is critical. CNN offers an efficient alternative, with 78 times fewer parameters and shorter training time.

This work provides empirical evidence of the effectiveness of unimodal models for classifying hate speech in Mexican Spanish and offers practical tools for content moderation systems to help create safer and more inclusive digital spaces.

Keywords: Unimodal Neural Networks, social media, misogyny

Capítulo 1

Generalidades

1.1. Introducción

La detección de mensajes misóginos en redes sociales ha cobrado relevancia en los últimos años debido al crecimiento exponencial de las plataformas digitales y la propagación de discursos de odio. Diversas investigaciones han abordado el problema desde distintas perspectivas, empleando múltiples enfoques metodológicos y tecnológicos, destacando el uso de redes neuronales y técnicas de aprendizaje profundo. (Amores et al., 2021) desarrollaron un detector de discurso de odio por ideología política en tuits en español, destacando la complejidad del lenguaje utilizado en las redes sociales y la necesidad de modelos que comprendan el contexto semántico de los mensajes. Aunque este estudio demuestra la viabilidad técnica de la detección automatizada en español, su enfoque se centra exclusivamente en el odio ideológico, sin abordar las características específicas del lenguaje misógino que requiere marcos de referencia distintos relacionados con género y violencia simbólica. Esta limitación evidencia la necesidad de modelos especializados que capturen las particularidades semánticas del discurso misógino, justificando el desarrollo de arquitecturas neuronales unimodales específicamente diseñadas para esta tarea.

De forma similar, (Andújar Carracedo, 2020) aplicó métodos de aprendizaje automático para detectar humor en Twitter, demostrando la utilidad de los modelos de clasificación en el tratamiento de datos textuales, que si bien este trabajo establece precedentes metodológicos valiosos para la clasificación de contenido en redes sociales, su enfoque en la detección de humor no considera las complejidades inherentes a la misoginia. Esta brecha metodológica subraya la importancia de

desarrollar modelos que no solo detectan patrones lingüísticos, sino que comprenden el contexto social y cultural específico de la misoginia, validando la necesidad de arquitecturas neuronales especializadas. (Aragón et al., 2018) en el marco del MEX-A3T de IberEval 2018 abordaron el análisis de autoría y agresividad en tuits en español, evidenciando la importancia de considerar las particularidades lingüísticas y culturales del español mexicano. Aunque este estudio reconoce la relevancia del contexto cultural hispanoamericano, su aproximación general a la “agresividad” no diferencia entre tipos específicos de discurso de odio, particularmente la misoginia, que presenta características lingüísticas y culturales distintivas. Por su parte, (Cerva Cerna, 2020) exploró la misoginia en el discurso institucional y las redes sociodigitales en México, resaltando la incidencia de mensajes misóginos en contextos de protesta feminista. Aunque este estudio proporciona evidencia crucial sobre la prevalencia y características de la misoginia digital en el contexto mexicano, carece de herramientas tecnológicas automatizadas para la detección y clasificación sistemática de estos mensajes. La brecha entre el análisis cualitativo exhaustivo y la falta de sistemas automatizados de detección hace presente la urgente necesidad de desarrollar modelos de redes neuronales unimodales que puedan escalar estos hallazgos cualitativos en sistemas de detección eficientes.

En cuanto a los modelos computacionales, (Badjatiya et al., 2017) emplearon técnicas de aprendizaje profundo para la detección de discurso de odio en tuits, utilizando redes neuronales recurrentes y embeddings de palabras. Este trabajo pionero estableció la viabilidad del aprendizaje profundo para la detección de odio, pero sus arquitecturas neuronales son relativamente básicas y no aprovechan los avances recientes en redes neuronales unimodales. Por lo que este enfoque fue ampliado por (Malik et al., 2023) y (Mathew et al., 2022), quienes realizaron estudios comparativos sobre la eficacia de distintas arquitecturas de redes neuronales para la detección de odio en redes sociales. Estos estudios sí proporcionan evaluaciones valiosas de diferentes arquitecturas, pero sus comparaciones se limitan a enfoques generales de la detección de odio sin explorar específicamente las ventajas de las redes neuronales unimodales para la clasificación de misoginia.

Particularmente en el ámbito de la misoginia, (Lagos, 2021) implementó clasificadores supervisados para la detección de mensajes misóginos en textos cortos, mientras que (López-Monroy et al., 2024) exploraron la transferencia de conocimiento desde letras de canciones para identificar

contenido misógino en redes sociales. (Plaza-Del-Arco et al., 2020) también contribuyeron con el uso de tecnologías del lenguaje para detectar misoginia y xenofobia en tuits españoles. Aunque (Lagos, 2021) establece un precedente directo para la detección automatizada de misoginia, su enfoque se limita a técnicas tradicionales de aprendizaje automático sin explorar las capacidades superiores de las redes neuronales profundas. Por otro lado, (López-Monroy et al., 2024) presenta una aproximación innovadora de transferencia de conocimiento, pero se limita a una fuente específica de datos (canciones) sin explorar la generalización a través de arquitecturas neuronales unimodales. Estas limitaciones metodológicas evidencian la oportunidad de desarrollar modelos que combinan la especialización en misoginia con las ventajas técnicas de las redes neuronales unimodales modernas. (García-Díaz et al., 2023) evaluaron estrategias de combinación de características utilizando transformadores y características lingüísticas, lo que ha permitido mejorar la detección de discurso de odio. Si bien este estudio avanza en la optimización de características para detección de odio, su enfoque en transformadores no explora específicamente las ventajas de las redes neuronales unimodales, que ofrecen beneficios únicos en términos de interpretabilidad y eficiencia computacional, representando una oportunidad para desarrollar estrategias de selección de características en redes neuronales unimodales para misoginia. Por otro lado, (Liu et al., 2024) revisaron los modelos de aprendizaje profundo aplicados a la detección de discurso de odio en redes sociales, destacando los avances en la interpretabilidad y robustez de estos modelos. Aunque esta revisión proporciona un panorama amplio del estado del arte, identifica una brecha significativa en la aplicación específica de redes neuronales unimodales para la detección de misoginia. Los autores reconocen la necesidad de modelos más interpretables y robustos, características que las redes neuronales unimodales pueden proporcionar de manera superior para tareas específicas como la clasificación de contenido misógino, justificando la propuesta de investigación.

Desde una perspectiva ética, (Aslam, 2024) abordó las consideraciones éticas en la detección de discurso de odio, subrayando los riesgos asociados con los sesgos de los modelos de inteligencia artificial y la importancia de desarrollar soluciones responsables.

La presente investigación se enfoca en el uso de redes neuronales unimodales para clasificar mensajes misóginos en redes sociales, aprovechando la capacidad de estos modelos para procesar datos textuales de forma eficiente. A diferencia de las aproximaciones multimodales, que

integran distintos tipos de datos, las redes unimodales ofrecen ventajas en la interpretabilidad y la optimización computacional, como lo destacan (Goyal y Cervone, 2023) en su análisis sobre las aplicaciones y limitaciones de estas arquitecturas.

Estos antecedentes permiten contextualizar la relevancia y los retos del tema, evidenciando la necesidad de soluciones precisas, eficientes y éticamente responsables para mitigar la propagación de mensajes misóginos en plataformas digitales.

1.2. Planteamiento del problema

1.2.1. Definición del problema

En la era digital actual, las redes sociales se han convertido en espacios clave para la interacción humana, permitiendo la difusión masiva de información y la libre expresión de opiniones. Sin embargo, esta apertura ha traído consigo la proliferación de mensajes de odio, entre ellos los de carácter misógino, que afectan gravemente la dignidad y el bienestar de las mujeres en los entornos digitales (Tortajada y Vera, 2021). La misoginia en redes sociales no solo perpetúa estereotipos de género y violencia simbólica, sino que también contribuye a la normalización de la discriminación y la violencia de género (CervaCerna2020).

A pesar de los esfuerzos realizados para mitigar este problema, la detección automática de mensajes misóginos sigue siendo un reto. La complejidad del lenguaje, el uso de ironía, sarcasmo y eufemismos dificultan la identificación precisa de estos contenidos (Amores et al., 2021; Lagos, 2021). Asimismo, las variaciones lingüísticas y culturales del español, particularmente en regiones como México, agravan el desafío de desarrollar modelos universales y eficaces para la clasificación de mensajes de odio (Aragón et al., 2018; López-Monroy et al., 2024).

Diversos enfoques han sido propuestos para abordar la detección de discurso de odio en redes sociales. Los modelos de aprendizaje profundo, como las redes neuronales recurrentes y las arquitecturas basadas en transformadores, han demostrado un rendimiento destacado en la clasificación de texto (Badjatiya et al., 2017; García-Díaz et al., 2023). Sin embargo, estas aproximaciones suelen ser multimodales, integrando datos de texto, imágenes y metadatos, lo que, si bien puede mejorar la precisión, incrementa la complejidad computacional y la necesidad

de recursos de procesamiento (Malik et al., 2023; Mathew et al., 2022).

En este contexto, las redes neuronales unimodales, centradas exclusivamente en datos textuales, emergen como una alternativa prometedora al ofrecer soluciones más eficientes y con menor demanda computacional (Goyal y Cervone, 2023; Zhang et al., 2024). Estas arquitecturas permiten un procesamiento más directo y especializado del texto, facilitando la interpretabilidad de los resultados, aspecto crucial en aplicaciones sensibles como la detección de mensajes misóginos.

Sin embargo, a pesar de sus ventajas, la literatura revela una brecha en la investigación enfocada específicamente en el uso de redes neuronales unimodales para la clasificación de mensajes misóginos en redes sociales. La mayoría de los estudios existentes se centran en el discurso de odio general (Amores et al., 2021; Shaip, 2024) o en la agresividad en línea (Aragón et al., 2018), sin abordar de manera particular las características lingüísticas y semánticas que distinguen la misoginia en las plataformas digitales.

Esta problemática se vuelve aún más relevante al considerar el impacto psicológico que la exposición constante a estos mensajes tiene en las usuarias de redes sociales, así como en los moderadores de contenido que deben enfrentarse diariamente a este tipo de materiales (Vasco Baracaldo, 2023). Además, la ineficacia de los sistemas actuales para identificar con precisión estos mensajes contribuye a la perpetuación del acoso y la discriminación en línea (Yang et al., 2021).

Aunque existe una cantidad creciente de estudios y documentos que abordan la detección del discurso de odio en redes sociales, son escasos aquellos que combinan el uso de redes neuronales con un enfoque específico en el análisis de contenido en español de México. Destaca la necesidad de desarrollar soluciones más precisas y adaptadas al contexto mexicano, lo que convierte este proyecto en una contribución valiosa a la implementación de las redes neuronales y al combate del discurso misógino hacia las mujeres en entornos digitales.

Por lo tanto, se plantea la necesidad de desarrollar un modelo de redes neuronales unimodales que permita la detección eficiente y precisa de mensajes misóginos en redes sociales, abordando las limitaciones de las aproximaciones existentes y contribuyendo a la creación de espacios digitales más seguros e inclusivos.

1.2.2. Delimitación de la investigación

Alcances

Se implementa un modelo de red neuronal unimodal enfocado exclusivamente en datos textuales para la clasificación de mensajes misóginos en redes sociales.

El enfoque se centra en mensajes escritos en español, lo que beneficia a comunidades hispanohablantes y permite abordar problemáticas específicas de esta lengua.

Los resultados se documentan y publican para servir como base para futuras investigaciones en el campo de la detección de discurso de odio y la aplicación de redes neuronales.

Limitaciones

La calidad del modelo depende en gran medida de la precisión y representatividad del dataset utilizado para su entrenamiento.

Al utilizar un enfoque unimodal, este proyecto se centra únicamente en texto, sin considerar información contextual adicional como metadatos, historial de publicaciones del usuario o interacciones en redes sociales, imágenes, audio o video.

El proyecto no incluye optimizaciones para su implementación en tiempo real o en sistemas de alto rendimiento.

1.3. Pregunta de investigación e hipótesis

¿Cómo pueden las redes neuronales unimodales clasificar de manera eficiente mensajes con contenido misógino en el contexto del español mexicano?

El uso de redes neuronales unimodales, optimizadas para la clasificación de texto en el lenguaje de español mexicano, permite detectar mensajes misóginos con una precisión **superior al 85 %**.

1.4. Objetivo general

Desarrollar un modelo basado en redes neuronales unimodales para la clasificación de mensajes misóginos en español mexicano.

1.5. Objetivos específicos

- Recopilar y seleccionar datasets representativos de mensajes en español mexicano que contengan ejemplos de lenguaje misógino y no misógino para el entrenamiento y evaluación del modelo.
- Diseñar y configurar una arquitectura de red neuronal unimodal específica para la detección de patrones lingüísticos asociados a la misoginia en el contexto del español mexicano.
- Analizar los resultados para identificar patrones lingüísticos específicos del español mexicano que caracterizan al lenguaje misógino.
- Evaluar la eficiencia del modelo en la detección de mensajes misóginos y realizar las adecuaciones que sean necesarias.

1.6. Justificación

La detección y clasificación de mensajes misóginos en redes sociales es un problema creciente que requiere soluciones tecnológicas eficaces y éticas. La misoginia en plataformas digitales no solo perpetúa la discriminación de género, sino que también contribuye a un entorno hostil que afecta la salud mental y el bienestar de las mujeres y otros grupos vulnerables (**Mujeres2023**; Vasco Baracaldo, 2023). Ante este panorama, el uso de redes neuronales unimodales para la clasificación de mensajes misóginos se presenta como una alternativa prometedora, dado su potencial para procesar grandes volúmenes de datos y su capacidad para aprender patrones complejos en textos cortos, como los tuits (Goyal y Cervone, 2023; Zhang et al., 2024).

La realización de este estudio se justifica por la necesidad de desarrollar soluciones que utilicen tecnologías de vanguardia para abordar esta problemática. En particular, el uso de redes neuronales unimodales centradas en datos textuales constituye una alternativa prometedora para la clasificación automática de mensajes misóginos. A diferencia de los modelos multimodales, los unimodales ofrecen ventajas en términos de simplicidad computacional, menor demanda de recursos y mayor interpretabilidad, lo que facilita su implementación en plataformas digitales con limitaciones técnicas (Goyal y Cervone, 2023; Zhang et al., 2024).

Desde una perspectiva social, la implementación de este modelo puede contribuir a la creación de espacios digitales más seguros e inclusivos, reduciendo la exposición a contenido dañino y favoreciendo la participación activa de las mujeres en las plataformas digitales. Desde el ámbito académico, la investigación aporta a la literatura existente en detección de discurso de odio, brindando resultados experimentales y metodológicos que pueden servir como base para estudios futuros.

Finalmente, este estudio busca contribuir al campo del procesamiento del lenguaje natural (PLN) y la inteligencia artificial (IA) al explorar el potencial de las redes neuronales unimodales en un contexto específico y socialmente relevante. Al hacerlo, se espera no solo avanzar en el estado del arte en la detección de misoginia en redes sociales, sino también generar un impacto positivo en la sociedad al reducir la propagación de discursos de odio y promover un uso más ético de la tecnología (Aslam, 2024; Davidson et al., 2019).

1.7. Metodología utilizada

Para el desarrollo de esta investigación se adoptó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que proporciona un marco estructurado y ampliamente reconocido para proyectos de ciencia de datos y aprendizaje automático. Esta metodología resulta especialmente apropiada para el presente estudio, ya que permite abordar de manera sistemática el problema de clasificación de mensajes misóginos mediante redes neuronales unimodales.

Fase 1: Comprensión del problema

Se partió del reconocimiento de una problemática social urgente: la proliferación de mensajes misóginos en redes sociales y su impacto en la participación de las mujeres en espacios digitales. En esta etapa inicial se definen claramente los objetivos de la investigación y se establecen los criterios de éxito. Se realizó una revisión exhaustiva de la literatura existente para comprender el estado actual de la detección automática de misoginia en redes sociales, identificando las limitaciones de los enfoques multimodales y las oportunidades que ofrecen las redes neuronales unimodales. Durante esta fase se preparó la pregunta de investigación, se delimitó el alcance del estudio al contexto del español mexicano, y se definieron las métricas de evaluación que determinarían el éxito del modelo propuesto. También se identificaron las consideraciones éticas

necesarias para el manejo responsable de contenido potencialmente ofensivo.

Fase 2: Comprensión de los datos

Se procedió a la exploración y análisis de los conjuntos de datos disponibles, específicamente MEX-A3T IberEval 2018 y EXIST IberLEF 2024. Esta etapa incluyó la caracterización detallada de las fuentes de datos, el análisis de la distribución de clases, la identificación de patrones lingüísticos relevantes y la evaluación de la calidad y representatividad de los datos. Se realizó un análisis estadístico descriptivo para comprender la naturaleza de los mensajes, la variabilidad del lenguaje utilizado y las características específicas del español mexicano presentes en los conjuntos de datos. Esta comprensión profunda de los datos resultó fundamental para las decisiones posteriores en el preprocesamiento y modelado.

Fase 3: Preparación de los datos

La preparación de datos constituyó una fase crítica, implementando técnicas específicas de limpieza y normalización adaptadas al contexto de redes sociales. Se desarrolló un pipeline de preprocesamiento que incluyó la eliminación de duplicados, conversión a minúsculas, eliminación de URL, menciones y caracteres especiales, así como la normalización de espacios en blanco. Durante esta etapa se tomaron decisiones metodológicas importantes sobre qué elementos conservar o eliminar del texto, considerando el equilibrio entre la limpieza necesaria para el procesamiento automático y la preservación de características lingüísticas relevantes para la detección de misoginia. Se implementaron técnicas de tokenización y vectorización apropiadas para el contexto unimodal.

Fase 4: Modelado

Se diseñaron y configuraron diferentes arquitecturas de redes neuronales unimodales, experimentando con distintos ajustes para optimizar el rendimiento en la tarea específica de clasificación binaria. Esta fase incluye la selección de algoritmos apropiados, la definición de la arquitectura de red y la configuración de hiperparámetros. Se implementó un proceso de validación cruzado para asegurar la robustez de los resultados, y se realizaron múltiples experimentos para identificar la configuración óptima. El modelado se orientó específicamente hacia la maximización de la precisión mientras se mantiene un equilibrio adecuado con otras métricas como sensibilidad y especificidad.

Fase 5: Evaluación

La evaluación del modelo se realizó utilizando métricas estándar en tareas de clasificación binaria, con especial énfasis en alcanzar el objetivo de superar el 85 % de precisión establecido en la hipótesis. Se implementó un conjunto de datos separados para entrenamiento, validación y prueba, y se evaluó no solo el rendimiento cuantitativo del modelo, sino también su capacidad de generalización a diferentes tipos de mensajes misóginos y su robustez ante variaciones en el lenguaje. Esta fase incluyó análisis de errores detallados para identificar patrones en las clasificaciones incorrectas y oportunidades de mejora. Esta evaluación permite validar la hipótesis planteada en la investigación y comparar los resultados con enfoques previos del estado del arte.

Fase 6: Implementación y validación

La fase final se enfoca en la validación práctica del modelo desarrollado y la documentación completa de los resultados. Se documentan las lecciones aprendidas, las limitaciones identificadas y las recomendaciones para investigaciones futuras. Esta fase también incluye la preparación de la documentación técnica necesaria para la reproducibilidad del estudio y la posible transferencia de conocimiento a otros contextos similares. La aplicación de la metodología CRISP-DM permite mantener un enfoque sistemático y riguroso a lo largo de toda la investigación, facilitando la trazabilidad de las decisiones tomadas y asegurando que cada etapa contribuirá efectivamente al logro de los objetivos planteados.

Capítulo 2

Marco teórico

2.1. Introducción

La misoginia se define como el odio, desprecio o prejuicio hacia las mujeres, manifestándose en actitudes y comportamientos que las denigran y subestiman, basados en estereotipos de género negativos. En el contexto digital, esta hostilidad se traduce en agresiones como insultos, amenazas y descalificaciones en plataformas en línea, perpetuando una cultura de violencia y exclusión hacia las mujeres. Según un informe de ONU Mujeres, la violencia digital contra las mujeres es una extensión de la violencia de género, facilitada por las tecnologías de la información y comunicación, afectando su participación en espacios digitales y limitando su libertad de expresión (Mujeres, 2023).

Las redes sociales son plataformas digitales que permiten la interacción y el intercambio de información entre personas, formando comunidades en línea basadas en intereses comunes. Estas herramientas han transformado la comunicación al eliminar barreras geográficas y temporales, influyendo en áreas como la interacción personal, la política y el activismo (Concepto.de, 2023). Sin embargo, este mismo potencial de conexión ha dado lugar a la proliferación de discursos de odio, incluyendo mensajes misóginos que afectan desproporcionadamente a las mujeres en estos espacios digitales.

La Inteligencia Artificial (IA) se refiere a la capacidad de los sistemas computacionales para realizar tareas que normalmente requieren inteligencia humana, como el reconocimiento de patrones, la toma de decisiones y el aprendizaje autónomo. Dentro de la IA, el aprendizaje profundo

(Deep Learning) es una subárea que emplea redes neuronales artificiales diseñadas para procesar grandes volúmenes de datos y extraer patrones complejos (IBM, 2024). Este enfoque ha revolucionado campos como el procesamiento del lenguaje natural, la visión por computadora y el análisis de datos, permitiendo avances significativos en tareas como la clasificación de texto, la traducción automática y la detección de anomalías. En el contexto de las redes sociales, estas tecnologías son esenciales para analizar el contenido generado por los usuarios, identificar discursos de odio y personalizar experiencias, destacando su papel como una herramienta clave para abordar problemas contemporáneos en el entorno digital (CulturaAI, 2023).

2.2. Conceptos y teorías fundamentales de la investigación

2.2.1. Redes neuronales artificiales

Las redes neuronales artificiales (RNA) son modelos computacionales inspirados en la estructura y funcionamiento del cerebro humano, diseñados para reconocer patrones y resolver problemas complejos mediante el aprendizaje a partir de datos. Una RNA está compuesta por unidades de procesamiento llamadas neuronas artificiales, organizadas en capas: una capa de entrada que recibe los datos, una o más capas ocultas que procesan la información, y una capa de salida que produce el resultado final (Goodfellow et al., 2016).

Cada neurona artificial realiza dos operaciones fundamentales. Primero, calcula una suma ponderada de sus entradas, donde cada entrada x_i se multiplica por un peso w_i y se añade un término de sesgo b :

$$z = \sum_{i=1}^n w_i x_i + b \quad (2.1)$$

Posteriormente, este resultado se transforma mediante una función de activación no lineal $f(z)$, que introduce la capacidad de modelar relaciones complejas. Entre las funciones de activación más utilizadas se encuentran:

- **Sigmoid:** $\sigma(z) = \frac{1}{1+e^{-z}}$, que mapea valores a un rango entre 0 y 1, útil para clasificación binaria.

- **ReLU** (Rectified Linear Unit): $f(z) = \max(0, z)$, que permite un entrenamiento más rápido y mitiga el problema del desvanecimiento del gradiente.
- **Softmax**: $f(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$, utilizada en la capa de salida para clasificación multiclase.

El proceso de entrenamiento de una red neuronal se basa en el algoritmo de retropropagación (backpropagation), que ajusta iterativamente los pesos de la red para minimizar una función de pérdida L que mide la diferencia entre las predicciones del modelo y los valores reales. Este ajuste se realiza mediante el descenso del gradiente:

$$w_{ij}^{nuevo} = w_{ij}^{actual} - \eta \frac{\partial L}{\partial w_{ij}} \quad (2.2)$$

donde η es la tasa de aprendizaje que controla la magnitud de los ajustes en cada iteración (Rumelhart et al., 1986).

2.2.2. Redes neuronales unimodales versus multimodales

Las redes neuronales pueden clasificarse según el tipo de datos que procesan. Las redes neuronales unimodales son modelos de aprendizaje profundo diseñados para trabajar con un solo tipo de dato, como texto, imágenes, audio o video. A diferencia de los enfoques multimodales, que integran múltiples modalidades de datos simultáneamente, estos modelos se enfocan únicamente en una modalidad, lo que los hace más simples y eficientes computacionalmente (Goyal y Cervone, 2023; Zhang et al., 2024).

En el contexto de la clasificación de texto, las redes unimodales ofrecen ventajas significativas:

- **Eficiencia computacional**: Requieren menos recursos de memoria y procesamiento al trabajar con un solo tipo de dato.
- **Interpretabilidad**: Facilitan la comprensión de qué características textuales influyen en las predicciones del modelo.
- **Especialización**: Permiten un ajuste fino específico para tareas de procesamiento de lenguaje natural.
- **Menor complejidad**: Simplifican el diseño, entrenamiento y mantenimiento del modelo.

Estas características son especialmente útiles en tareas como la detección de discurso de odio, donde el análisis del contenido textual es suficiente para identificar patrones complejos y resolver problemas específicos sin necesidad de integrar información visual o de audio (Badjatiya et al., 2017).

2.2.3. Arquitecturas de redes neuronales para procesamiento de texto

2.2.3.1. Redes Neuronales Recurrentes (RNN) y LSTM

Las Redes Neuronales Recurrentes (RNN) son arquitecturas diseñadas específicamente para procesar secuencias de datos, como texto o series temporales. A diferencia de las redes feedforward tradicionales, las RNN mantienen un estado oculto que captura información de elementos anteriores en la secuencia, permitiendo modelar dependencias temporales (Elman, 1990).

Sin embargo, las RNN tradicionales sufren del problema del desvanecimiento del gradiente, lo que dificulta el aprendizaje de dependencias a largo plazo. Las redes Long Short-Term Memory (LSTM) fueron diseñadas para resolver esta limitación mediante la introducción de un mecanismo de puertas que regula el flujo de información (Hochreiter y Schmidhuber, 1997).

Una unidad LSTM consta de tres puertas principales:

- **Puerta de olvido** (f_t): Determina qué información del estado de celda anterior debe descartarse.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2.3)$$

- **Puerta de entrada** (i_t): Decide qué nueva información se almacenará en el estado de celda.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2.4)$$

- **Puerta de salida** (o_t): Controla qué parte del estado de celda se utiliza para calcular la salida.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (2.5)$$

El estado de celda C_t se actualiza combinando la información olvidada y la nueva información:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (2.6)$$

donde $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$ es el candidato de nuevo estado de celda, y $\odot$ representa el producto elemento a elemento.

En tareas de clasificación de texto, las LSTM han demostrado ser efectivas para capturar dependencias contextuales en secuencias de palabras, especialmente en textos cortos como tweets (Badjatiya et al., 2017).

2.2.3.2. Redes Neuronales Convolucionales para texto

Aunque originalmente diseñadas para procesamiento de imágenes, las Redes Neuronales Convolucionales (CNN) han demostrado ser altamente efectivas en tareas de clasificación de texto (Kim, 2014). La clave de su éxito radica en su capacidad para detectar patrones locales relevantes, como n-gramas significativos, mediante operaciones de convolución.

En el procesamiento de texto, una CNN opera sobre una matriz de embeddings de palabras. Una capa convolucional aplica múltiples filtros que se deslizan sobre la secuencia de palabras, capturando características locales:

$$c_i = f(w \cdot x_{i:i+h-1} + b) \quad (2.7)$$

donde $x_{i:i+h-1}$ representa una ventana de h palabras consecutivas, w es el vector de pesos del filtro, b es el término de sesgo, y f es una función de activación no lineal (típicamente ReLU).

Tras la convolución, se aplica una operación de max-pooling que extrae la característica más importante de cada mapa de características:

$$\hat{c} = \max(c_1, c_2, \dots, c_{n-h+1}) \quad (2.8)$$

Esta arquitectura permite capturar automáticamente patrones lingüísticos relevantes para la clasificación, como expresiones idiomáticas o combinaciones de palabras asociadas con misoginia, sin necesidad de ingeniería manual de características (Zhang et al., 2024).

2.2.3.3. Modelos basados en Transformers

La arquitectura Transformer, introducida por Vaswani et al. (2017), revolucionó el procesamiento del lenguaje natural al proponer un mecanismo de atención que prescinde completamente de la recurrencia. El componente central es el mecanismo de auto-atención (self-attention), que permite a cada palabra en una secuencia atender a todas las demás palabras, capturando dependencias contextuales complejas independientemente de la distancia entre ellas.

El mecanismo de atención calcula una representación ponderada de las palabras en una secuencia mediante tres componentes: consultas (queries), claves (keys) y valores (values). Para cada palabra, la atención se calcula como:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.9)$$

donde Q , K y V son matrices de consultas, claves y valores respectivamente, y d_k es la dimensionalidad de las claves, utilizada para escalar el producto punto.

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) llevó esta arquitectura un paso más allá al introducir el preentrenamiento bidireccional. BERT se entrena mediante dos tareas:

- **Masked Language Modeling (MLM)**: Se enmascara aleatoriamente el 15 % de las palabras en una oración y el modelo debe predecirlas basándose en el contexto bidireccional.
- **Next Sentence Prediction (NSP)**: El modelo aprende a determinar si dos oraciones son consecutivas en un texto.

Para el español, BETO (Cañete et al., 2020) es una adaptación de BERT entrenada en grandes corpus de texto en español (aproximadamente 3 mil millones de palabras de Wikipedia y otros recursos). BETO utiliza el tokenizador WordPiece, que divide las palabras en subpalabras, permitiendo manejar vocabulario abierto y palabras fuera del vocabulario de manera efectiva. Esta especialización lingüística es crucial para capturar las particularidades morfológicas, sintácticas y semánticas del español, incluyendo variantes dialectales como el español mexicano.

En tareas de clasificación de texto, BETO se fine-tunea añadiendo una capa de clasificación sobre la representación del token especial [CLS]:

$$y = \text{softmax}(W \cdot h_{[CLS]} + b) \quad (2.10)$$

donde $h_{[CLS]}$ es la representación contextual del token [CLS] producida por BETO, y W y b son parámetros entrenables de la capa de clasificación.

2.2.4. Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN) es una disciplina de la inteligencia artificial que se enfoca en la interacción entre computadoras y lenguaje humano. En el contexto de la detección de misoginia, el PLN proporciona las herramientas necesarias para transformar texto no estructurado en representaciones numéricas que los modelos de aprendizaje profundo pueden procesar (Manning y Schütze, 1999).

2.2.4.1. Preprocesamiento de texto

El preprocesamiento es una etapa fundamental que transforma el texto crudo en un formato adecuado para el análisis computacional. Las técnicas comunes incluyen:

- **Tokenización:** Segmentación del texto en unidades lingüísticas (palabras, subpalabras o caracteres).
- **Normalización:** Conversión a minúsculas, eliminación de acentos o caracteres especiales.
- **Eliminación de ruido:** Remoción de URLs, menciones, hashtags y otros elementos no informativos típicos de redes sociales.
- **Lematización o stemming:** Reducción de palabras a su forma base o raíz para reducir la variabilidad léxica.

En el contexto de redes sociales, el preprocesamiento debe balancear la eliminación de ruido con la preservación de características lingüísticas relevantes para la detección de misoginia, como el uso intencional de mayúsculas o signos de puntuación para expresar énfasis o emoción (Aragón et al., 2018).

2.2.4.2. Representaciones de palabras: Embeddings

Los embeddings son representaciones vectoriales densas de palabras en un espacio continuo de baja dimensionalidad, donde palabras con significados similares se ubican cerca entre sí. A diferencia de las representaciones one-hot tradicionales, los embeddings capturan relaciones semánticas y sintácticas entre palabras (Mikolov et al., 2013).

Existen dos paradigmas principales de embeddings:

1. **Embeddings estáticos:** Modelos como Word2Vec (Mikolov et al., 2013) y GloVe (Pennington et al., 2014) asignan una representación fija a cada palabra, independiente del contexto en que aparece.
2. **Embeddings contextuales:** Modelos como BERT generan representaciones dinámicas que varían según el contexto, capturando fenómenos como la polisemia. Por ejemplo, la palabra “banco” tendrá representaciones diferentes en “banco de plaza” versus “banco financiero”.

Los embeddings contextuales han demostrado ser superiores en tareas de clasificación de texto, particularmente cuando el significado de las palabras depende fuertemente del contexto, como ocurre con el lenguaje sarcástico o irónico frecuente en discursos misóginos (Devlin et al., 2019).

2.2.5. Métricas de evaluación en clasificación de texto

La evaluación rigurosa de modelos de clasificación requiere métricas que capturen diferentes aspectos del rendimiento. En problemas de clasificación binaria como la detección de misoginia, las métricas fundamentales se derivan de la matriz de confusión, que compara las predicciones del modelo con las etiquetas reales:

- **Verdaderos Positivos (TP):** Mensajes misóginos correctamente identificados.
- **Verdaderos Negativos (TN):** Mensajes no misóginos correctamente identificados.
- **Falsos Positivos (FP):** Mensajes no misóginos incorrectamente clasificados como misóginos.

- **Falsos Negativos (FN):** Mensajes misóginos incorrectamente clasificados como no misóginos.

A partir de estos valores, se calculan las siguientes métricas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.11)$$

$$Precision = \frac{TP}{TP + FP} \quad (2.12)$$

$$Recall(Sensibilidad) = \frac{TP}{TP + FN} \quad (2.13)$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (2.14)$$

El F1-Score es particularmente relevante cuando existe desbalance de clases, ya que proporciona una media armónica entre precisión y recall. Adicionalmente, el Macro-F1, que promedia el F1-Score de cada clase de manera independiente, es recomendado en competencias como IberEval e IberLEF para evaluar el rendimiento global sin sesgo hacia la clase mayoritaria (Fersini et al., 2018).

2.2.6. Modelos de clasificación de discurso de odio

Los modelos de clasificación de discurso de odio son herramientas basadas en aprendizaje automático que identifican y categorizan mensajes ofensivos en plataformas digitales. Utilizan técnicas de procesamiento del lenguaje natural (PLN) y redes neuronales para analizar el contexto y los patrones lingüísticos en textos etiquetados. Estos modelos son útiles para moderar contenido y estudiar tendencias de odio, aunque enfrentan retos como la detección de sarcasmo, ironía y los sesgos en los datos de entrenamiento (Liu et al., 2024).

La evolución de estos modelos ha progresado desde enfoques basados en características lingüísticas manuales (n-gramas, léxicos de palabras ofensivas) hasta modelos de aprendizaje profundo que aprenden automáticamente representaciones jerárquicas del texto. Los desafíos

específicos incluyen:

- **Ambigüedad contextual:** Palabras que pueden ser ofensivas en un contexto pero neutras en otro.
- **Evolución del lenguaje:** Nuevas expresiones y eufemismos que evaden la detección.
- **Variación dialectal:** Diferencias regionales en el uso del lenguaje, especialmente relevantes en español.
- **Desbalance de clases:** Predominancia de contenido no ofensivo sobre ofensivo en datasets reales.

La inteligencia artificial aplicada a la detección de discurso de odio tiene un impacto social significativo al contribuir a la creación de entornos digitales más seguros e inclusivos. Estas tecnologías ayudan a identificar y moderar mensajes ofensivos en tiempo real, protegiendo a los usuarios de contenido dañino. Además, permiten analizar tendencias de odio en redes sociales, proporcionando datos valiosos para crear políticas públicas y estrategias de prevención. Sin embargo, también generan desafíos éticos, como la privacidad de los usuarios, el riesgo de censura excesiva y los posibles sesgos en los modelos, que deben ser gestionados para garantizar un uso responsable de la IA (Aslam, 2024; Malik et al., 2023).

2.3. Literatura relacionada

2.3.1. Trabajos fundacionales en detección de discurso de odio

El campo de la detección automática de discurso de odio ha experimentado un desarrollo significativo en la última década. Badjatiya et al. (2017) establecieron las bases metodológicas al emplear técnicas de aprendizaje profundo para la detección de discurso de odio en Twitter, utilizando redes neuronales recurrentes y embeddings de palabras. Su trabajo demostró que los modelos de aprendizaje profundo superan consistentemente a los enfoques tradicionales basados en características manuales, logrando accuracy entre 91-93 % en datasets en inglés. Davidson et al. (2017) realizaron una contribución importante al distinguir entre discurso de odio y lenguaje

ofensivo no discriminatorio, reconociendo que no todo contenido ofensivo constituye discurso de odio dirigido a grupos protegidos. Esta distinción es fundamental para evitar la censura excesiva y mejorar la especificidad de los sistemas de detección.

Más recientemente, Liu et al. (2024) realizaron una revisión exhaustiva de los modelos de aprendizaje profundo aplicados a la detección de discurso de odio en redes sociales, destacando los avances en interpretabilidad y robustez de estos modelos. Su análisis identifica una brecha significativa en la aplicación específica de redes neuronales unimodales para la detección de misoginia, señalando la necesidad de modelos más interpretables y robustos, características que las arquitecturas unimodales pueden proporcionar de manera superior.

2.3.2. Investigaciones en detección de misoginia

La detección específica de contenido misógino ha emergido como un área de investigación con características distintivas respecto al discurso de odio general. Fersini et al. (2018) organizaron la primera tarea compartida de identificación de misoginia en Twitter, estableciendo benchmarks y metodologías que han guiado investigaciones posteriores. Sus hallazgos revelaron que la misoginia presenta patrones lingüísticos específicos que requieren modelos especializados. Lagos (2021) implementó clasificadores supervisados para la detección de mensajes misóginos en textos cortos, evaluando el desempeño de algoritmos como Support Vector Machines (SVM) y Naive Bayes. Los resultados demostraron que los modelos supervisados son efectivos para identificar patrones lingüísticos asociados a mensajes misóginos, aunque enfrentan limitaciones en el manejo del sarcasmo y la evolución del lenguaje. Este trabajo subraya la importancia de desarrollar herramientas automatizadas para combatir el contenido ofensivo en línea, sirviendo como una base para futuras investigaciones en contextos más específicos. López-Monroy et al. (2024) exploraron la transferencia de conocimiento desde letras de canciones para identificar contenido misógino en redes sociales, demostrando que el conocimiento de dominios relacionados puede mejorar el rendimiento en la detección de misoginia. Este enfoque innovador sugiere que los patrones misóginos trascienden plataformas específicas y pueden ser aprendidos de múltiples fuentes textuales.

2.3.3. Detección de discurso de odio en español

El desarrollo de modelos específicos para el español ha cobrado importancia debido a las particularidades lingüísticas y culturales del idioma. Aragón et al. (2018) en el marco del MEX-A3T de IberEval 2018 abordaron el análisis de autoría y agresividad en tuits en español, evidenciando la importancia de considerar las particularidades lingüísticas y culturales del español mexicano. Su dataset ha sido fundamental para el desarrollo de modelos contextualizados a variantes del español americano. Plaza-Del-Arco et al. (2020) contribuyeron con el uso de tecnologías del lenguaje para detectar misoginia y xenofobia en tuits españoles, abordando los retos específicos del idioma como el uso coloquial y las variaciones regionales, proponiendo soluciones adaptadas al contexto hispano. Su trabajo demostró que las características lingüísticas del español, como la flexión verbal rica y el orden de palabras flexible, requieren adaptaciones específicas en los modelos de detección.

En el ámbito de sistemas avanzados, García-Díaz et al. (2023) desarrollaron un sistema que combina el análisis de características lingüísticas con modelos de lenguaje basados en transformadores, alcanzando una precisión cercana al 90 % en la identificación de mensajes ofensivos en español. La herramienta extrae 365 rasgos lingüísticos, incluyendo aspectos fonéticos, morfosintácticos, semánticos y pragmáticos, lo que permite una comprensión profunda del contenido y contexto de los textos analizados. Este enfoque integral facilita la detección eficiente de discursos de odio relacionados con raza, género, orientación sexual, nacionalidad o religión. Plaza-del-Arco et al. (2024) en el contexto de EXIST (sEXism Identification in Social neTworks) enfatizan la importancia de la calidad de los datos y la necesidad de modelos adaptativos para abordar los retos de la detección de sexismo en redes sociales. Su trabajo proporciona benchmarks actualizados y datasets anotados con múltiples dimensiones del sexismo, incluyendo el tipo de discurso, la intención del autor y categorías específicas de violencia de género.

2.3.4. Aplicaciones de redes neuronales en clasificación de texto en español

Diversos trabajos han explorado la aplicación de arquitecturas neuronales específicas para tareas de clasificación en español. García Pérez et al. (2018) implementó una red neuronal artificial tipo perceptrón multicapa para clasificar tweets en tres categorías: situaciones de riesgo,

servicio de agua y política. Este enfoque demostró ser efectivo para tareas de clasificación multi-clase, superando en precisión a métodos como el clasificador bayesiano. Este trabajo subraya la capacidad de las redes neuronales para manejar datos lingüísticos complejos en español y para adaptarse a contextos específicos como el estado de Tamaulipas, México. Andújar Carracedo (2020) desarrolló técnicas de procesamiento del lenguaje natural (PLN) y aprendizaje automático para detectar humor en Twitter, demostrando la versatilidad de estas tecnologías para analizar el lenguaje humano en distintos contextos. Aunque enfocado en humor, este trabajo evidencia la aplicabilidad de métodos de PLN para capturar fenómenos lingüísticos sutiles que requieren comprensión contextual. González-Guerra et al. (2021) señalaron que las redes neuronales unimodales, como las Long Short-Term Memory (LSTM) y los modelos de atención, siguen siendo herramientas útiles en entornos donde los datos y los recursos computacionales son limitados, ofreciendo un balance entre simplicidad y efectividad en comparación con modelos multimodales más complejos. Esta versatilidad ha permitido su aplicación en tareas relacionadas con la clasificación de emociones y discursos tóxicos en plataformas sociales.

2.3.5. Modelos de lenguaje preentrenados para español

El desarrollo de BERT (Devlin et al., 2019) marcó un hito en el procesamiento del lenguaje natural, y su adaptación al español ha sido crucial para mejorar el rendimiento en tareas de PLN en este idioma. Cañete et al. (2020) desarrollaron BETO, una versión de BERT entrenada específicamente en corpus de español, incluyendo Wikipedia en español y corpus de noticias. BETO ha demostrado superar consistentemente a modelos multilingües en tareas de clasificación de texto en español, capturando mejor las particularidades morfosintácticas y semánticas del idioma.

La importancia de los modelos especializados en español se evidencia en los resultados comparativos reportados en la literatura, donde BETO típicamente supera a mBERT (multilingual BERT) por 3-8 puntos porcentuales en diversas tareas de PLN en español (Cañete et al., 2020). Esta ventaja es particularmente notable en tareas que requieren comprensión profunda del contexto cultural y lingüístico, como la detección de misoginia.

2.3.6. Desafíos y consideraciones éticas

Chakraborty et al. (2019) analizaron técnicas de aprendizaje automático aplicadas a la detección de contenido abusivo en textos, abordando métodos tradicionales y avanzados, como el aprendizaje profundo, para la extracción de características y clasificación. Su trabajo destaca los desafíos en redes sociales, como la variabilidad lingüística y la representación dispersa de textos, enfatizando la necesidad de datasets robustos y técnicas de preprocesamiento adaptadas.

Desde una perspectiva ética, Aslam (2024) abordó las consideraciones éticas en la detección de discurso de odio, subrayando los riesgos asociados con los sesgos de los modelos de inteligencia artificial y la importancia de desarrollar soluciones responsables. Los autores identifican tres desafíos éticos principales: (1) el balance entre libertad de expresión y protección contra el odio, (2) la transparencia y explicabilidad de las decisiones de los modelos, y (3) la representatividad de los datos de entrenamiento para evitar sesgos demográficos o culturales. Vasco Baracaldo (2023) estudiaron el impacto psicológico que la exposición constante a mensajes de odio tiene en las usuarias de redes sociales y en los moderadores de contenido que deben enfrentarse diariamente a este tipo de materiales. Sus hallazgos subrayan la urgencia de desarrollar sistemas automatizados efectivos que reduzcan la exposición humana a contenido dañino.

2.3.7. Perspectiva sociológica y cultural

Desde una perspectiva sociológica, Cerva Cerna (2020) exploró la misoginia en el discurso institucional y las redes sociodigitales en México, resaltando la incidencia de mensajes misóginos en contextos de protesta feminista. Este estudio proporciona evidencia crucial sobre la prevalencia y características de la misoginia digital en el contexto mexicano, identificando patrones específicos de violencia verbal que deben ser considerados en el desarrollo de sistemas de detección automatizada. Tortajada y Vera (2021) analizaron cómo la misoginia en redes sociales no solo perpetúa estereotipos de género y violencia simbólica, sino que también contribuye a la normalización de la discriminación y la violencia de género en espacios digitales y físicos. Su trabajo enfatiza la necesidad de intervenciones tecnológicas como parte de una estrategia más amplia de prevención de la violencia de género.

2.3.8. Análisis comparativo del estado del arte

Jahan y Oussalah (2023) realizaron una revisión sistemática que muestra que en los últimos años ha aumentado significativamente la cantidad de investigaciones sobre detección de lenguaje de odio. Su metaanálisis resalta que modelos avanzados de inteligencia artificial, como BERT y sus variantes, representan una oportunidad importante para mejorar esta tarea. Sin embargo, el principal problema identificado es la falta de datasets suficientes que incluyan diferentes culturas e idiomas, particularmente para variantes dialectales del español. Malik et al. (2023) y Mathew et al. (2022) realizaron estudios comparativos sobre la eficacia de distintas arquitecturas de redes neuronales para la detección de odio en redes sociales. Sus comparaciones sistemáticas revelan que mientras los transformers preentrenados generalmente superan a arquitecturas tradicionales como LSTM y CNN, estas últimas mantienen relevancia en escenarios con recursos limitados o cuando se requiere inferencia en tiempo real.

La Tabla 2.1 presenta una síntesis comparativa de los trabajos más relevantes en detección de misoginia y discurso de odio en español:

Tabla 2.1. Comparación de trabajos previos en detección de misoginia y discurso de odio

Autor(es)	Año	Arquitectura	Dataset	Idioma	Accuracy
Badjatiya et al.	2017	LSTM + GRU	Twitter (inglés)	Inglés	93 %
Fersini et al.	2018	Ensemble ML	AMI Task	Inglés/Español	70-75 %
Plaza et al.	2020	SVM + Features	Twitter ES	Español	75 %
Lagos	2021	SVM + NB	Textos cortos	Español	78 %
García-Díaz et al.	2023	Transformers + Features	Multiple	Español	90 %
Plaza et al.	2024	BERT variants	EXIST 2024	Español	82-85 %
Este trabajo	2025	BETO (Transformer)	AMI + EXIST	Español MX	88.42 %

Esta revisión del estado del arte revela varias brechas de investigación que este trabajo aborda:

1. **Especialización en español mexicano:** Mientras que la mayoría de trabajos utilizan español peninsular o mezclas dialectales, este estudio se enfoca específicamente en variantes mexicanas.
2. **Arquitecturas unimodales modernas:** Existe investigación limitada sobre la aplicación de transformers unimodales especializados en español (BETO) para detección de misoginia.

3. **Fusión de datasets:** La combinación de múltiples fuentes de datos etiquetados (AMI + EXIST) para aumentar la robustez y generalización de los modelos es una contribución metodológica importante.
4. **Análisis comparativo riguroso:** Pocos trabajos comparan sistemáticamente arquitecturas tradicionales (LSTM, CNN) con transformers modernos en la misma tarea y dataset.

2.4. Marco tecnológico

Para el desarrollo e implementación de este trabajo de investigación, se utilizaron las siguientes herramientas tecnológicas, seleccionadas por su madurez, documentación, soporte comunitario y eficacia demostrada en tareas de procesamiento del lenguaje natural:

2.4.1. Software y frameworks

Python 3.10: Lenguaje de programación principal, seleccionado por su amplia adopción en la comunidad de ciencia de datos e inteligencia artificial. Python ofrece un ecosistema rico de bibliotecas especializadas y una sintaxis que facilita la experimentación rápida y el desarrollo iterativo de modelos de aprendizaje profundo.

PyTorch 2.0: Framework de aprendizaje profundo desarrollado por Meta AI, seleccionado por su flexibilidad en la definición de arquitecturas personalizadas, su eficiente manejo de operaciones tensoriales mediante aceleración por GPU, y su paradigma de diferenciación automática que facilita la implementación de nuevas arquitecturas. PyTorch se ha consolidado como el framework preferido en investigación académica debido a su naturaleza más pythonica y su modo eager execution que simplifica el debugging (Paszke et al., 2019).

Hugging Face Transformers 4.30: Biblioteca que proporciona implementaciones preentrenadas de modelos basados en arquitecturas Transformer, incluyendo BERT, RoBERTa, y específicamente BETO para español. Esta biblioteca estandariza la interfaz para cargar, fine-tunar y utilizar modelos de lenguaje preentrenados, reduciendo significativamente el tiempo de desarrollo y garantizando implementaciones optimizadas (Wolf et al., 2020).

Scikit-learn 1.3: Biblioteca de aprendizaje automático que proporciona implementaciones efi-

cientes de métricas de evaluación, validación cruzada y técnicas de preprocesamiento. Utilizada principalmente para el cálculo de métricas de rendimiento (accuracy, precision, recall, F1-score) y la generación de matrices de confusión.

Pandas 2.0 y NumPy 1.24: Bibliotecas fundamentales para manipulación eficiente de datos estructurados y operaciones numéricas. Pandas facilita la carga, limpieza y transformación de datasets en formato tabular, mientras que NumPy proporciona estructuras de datos y operaciones de álgebra lineal de bajo nivel.

2.4.2. Recursos lingüísticos y datasets

MEX-A3T dataset (IberEval 2018): Conjunto de datos de 3,307 tweets anotados en español mexicano, desarrollado para el análisis de agresividad y misoginia en Twitter (Aragón et al., 2018). Este dataset incluye etiquetas binarias de misoginia y categorías específicas (estereotipos, dominación, acoso sexual, etc.), proporcionando ground truth para el entrenamiento supervisado.

EXIST dataset (IberLEF 2024): Dataset anotado por múltiples evaluadores que contiene aproximadamente 3,660 tweets en español con etiquetas de sexismo, tipo de discurso e información demográfica de los anotadores (Plaza et al., 2024). La anotación múltiple permite estimar la ambigüedad inherente en la clasificación de contenido misógino.

BETO (Spanish BERT): Modelo de lenguaje basado en la arquitectura BERT, preentrenado en aproximadamente 3 mil millones de palabras de texto en español de Wikipedia y otros corpus (Cañete et al., 2020). BETO utiliza tokenización WordPiece con un vocabulario de 31,002 tokens específicos del español, capturando las particularidades morfológicas y sintácticas del idioma. La versión utilizada es dccuchile/bert-base-spanish-wwm-cased (whole word masking), que mejora el preentrenamiento al enmascarar palabras completas en lugar de subpalabras.

SpaCy 3.5: Biblioteca de procesamiento del lenguaje natural utilizada para tareas auxiliares de preprocesamiento como tokenización avanzada, reconocimiento de entidades y análisis morfosintáctico cuando se requiere análisis lingüístico complementario.

2.4.3. Infraestructura computacional

Google Colaboratory: Entorno de desarrollo basado en Jupyter Notebooks con acceso gratuito a GPUs NVIDIA Tesla T4 (16GB de VRAM). Colab fue seleccionado por su accesibilidad, integración con Google Drive para almacenamiento de datasets y modelos, y su capacidad de proporcionar sesiones de GPU sin costo que permiten el entrenamiento de modelos basados en transformers. Las sesiones típicamente proporcionan 12-16 horas de tiempo de GPU continuo, suficiente para el fine-tuning de BETO.

MLflow 2.5: Plataforma de código abierto para gestión del ciclo de vida de modelos de aprendizaje automático. MLflow se utilizó para el seguimiento de experimentos, registro de hiperparámetros, métricas de evaluación y versionado de modelos. Esta herramienta facilitó la comparación sistemática entre diferentes configuraciones de arquitecturas y la reproducibilidad de los experimentos.

Weights & Biases (wandb): Plataforma complementaria para visualización y monitoreo de experimentos en tiempo real. Utilizada para rastrear curvas de aprendizaje, visualizar métricas durante el entrenamiento y comparar runs experimentales de manera interactiva.

2.4.4. Hardware local

Equipo de cómputo personal: Equipado con procesador Intel Core i5 de 12^a Generación y 16 GB de RAM DDR4. Utilizado para preprocesamiento de datos, análisis exploratorio, desarrollo de código y experimentación con modelos pequeños. Aunque insuficiente para el entrenamiento de transformers a escala completa, este equipo permitió el desarrollo y debugging de pipelines de datos y la validación de modelos más ligeros como LSTM y CNN.

2.4.5. Herramientas de control de versiones y documentación

Git y GitHub: Sistema de control de versiones utilizado para mantener el historial del código fuente, facilitar la colaboración y garantizar la reproducibilidad del trabajo. El repositorio incluye scripts de entrenamiento, notebooks de análisis y documentación técnica.

LaTeX: Sistema de composición de documentos utilizado para la redacción de esta tesis, seleccionado por su capacidad para manejar ecuaciones matemáticas complejas, referencias

bibliográficas y estructura de documentos académicos de manera profesional.

2.4.6. Justificación de la selección tecnológica

Esta infraestructura tecnológica se ha seleccionado considerando los siguientes criterios:

1. **Eficacia comprobada:** Todas las herramientas han demostrado efectividad en proyectos de PLN y detección de discurso de odio en trabajos previos.
2. **Escalabilidad:** PyTorch y la infraestructura de Google Colab permiten escalar el entrenamiento de modelos pequeños a arquitecturas basadas en transformers con millones de parámetros.
3. **Reproducibilidad:** El uso de MLflow, control de versiones y entornos virtuales garantiza que los experimentos sean reproducibles y verificables.
4. **Especialización lingüística:** BETO proporciona conocimiento preentrenado específico del español, crucial para capturar las particularidades del español mexicano.
5. **Accesibilidad:** La mayoría de las herramientas son de código abierto y gratuitas, facilitando la replicación del trabajo por otros investigadores.
6. **Soporte comunitario:** Todas las tecnologías cuentan con comunidades activas, documentación extensa y recursos de aprendizaje, facilitando la resolución de problemas técnicos.

Esta combinación de software, datos y hardware representa el estado actual de las mejores prácticas en investigación de procesamiento del lenguaje natural aplicado a la detección de discurso de odio, proporcionando una base sólida y confiable para el desarrollo de modelos de clasificación de contenido misógino en redes sociales.

Capítulo 3

Análisis exploratorio de datos

3.1. Introducción

El análisis exploratorio de datos es una etapa fundamental en el desarrollo de modelos de aprendizaje automático, ya que permite comprender las características del corpus, identificar patrones relevantes y tomar decisiones informadas sobre el preprocesamiento y la arquitectura del modelo. En este capítulo se describe el proceso de recopilación, exploración y preparación de los datos utilizados para entrenar y evaluar los modelos de clasificación de mensajes misóginos.

3.2. Recopilación de datos

Para el desarrollo de este estudio se utilizaron dos conjuntos de datos en lenguaje español mexicano reconocidos en tareas de detección de misoginia en redes sociales. La selección de estos datasets se fundamenta en su relevancia para el contexto mexicano, la calidad de sus anotaciones y su uso como benchmarks en competencias internacionales de procesamiento del lenguaje natural.

3.2.1. Dataset MEX-A3T

El dataset MEX-A3T fue desarrollado en el marco de IberEval 2018, una competición enfocada en la identificación automática de misoginia en tuits en español (Fersini et al., 2018). Este conjunto de datos contiene 3,307 tweets anotados manualmente como misóginos o no misóginos.

Una característica distintiva de este dataset es que incluye categorías específicas de misoginia como cosificación, dominación y estereotipos, lo que permite un análisis más detallado del tipo de violencia verbal ejercida.

Los datos proporcionados están separados por tabulaciones e incluyen los siguientes campos: `id`, `tweet`, `misogynous`, `misogyny_category` y `target`, donde:

- `id`: Denota un identificador único del tuit, utilizado para rastrear cada mensaje a lo largo del pipeline de procesamiento.
- `tweet`: Representa el texto completo del tuit, incluyendo menciones, hashtags, URLs y emojis en su formato original.
- `misogynous`: Define si el tuit es misógino o no mediante valores binarios. Toma el valor 1 si el tuit contiene contenido misógino y 0 si el tuit no presenta características misóginas.
- `misogyny_category`: Denota el tipo específico de comportamiento misógino presente en el mensaje. Este campo toma los siguientes valores:
 - `stereotype`: Denota la categoría "Stereotype & Objectification" (Estereotipos y Objetivación).
 - `dominance`: Indica la categoría "Dominance" (Dominación).
 - `derailing`: Indica la categoría "Derailing" (Descarrilamiento).
 - `sexual_harassment`: Indica la categoría "Sexual Harassment & Threats of Violence" (Acoso sexual y amenazas de violencia).
 - `discredit`: Indica la categoría "Discredit" (Descrédito).
 - 0: Si el tuit no es misógino.
- `target`: Indica el sujeto del tuit misógino, especificando si el ataque es dirigido o genérico. Toma los siguientes valores:
 - `active`: Indica un objetivo específico (individual), es decir, cuando el mensaje se dirige a una mujer o grupo de mujeres identificables.

- `passive`: Indica receptores potenciales (genérico), cuando el mensaje contiene misoginia pero no se dirige a un objetivo específico.
- `0`: Si el tuit no es misógino.

3.2.2. Dataset EXIST 2024

IberLEF (Iberian Languages Evaluation Forum) es una campaña de evaluación compartida de sistemas de Procesamiento del Lenguaje Natural en español y otras lenguas ibéricas que se organiza anualmente desde 2019. Su objetivo es promover la investigación en tareas de procesamiento, comprensión y generación de textos en español, proporcionando datasets estandarizados y métricas de evaluación comparables.

El dataset EXIST 2024 (sEXism Identification in Social neTworks) representa la última iteración de esta competición, anotado por múltiples evaluadores para garantizar la confiabilidad de las etiquetas. Contiene aproximadamente 3,660 tweets en español con etiquetas asociadas a la presencia de sexismo, tipo de discurso (por ejemplo, directo, reportado, juicio) y subcategorías específicas como violencia sexual o simbólica. Una característica distintiva de este dataset es que incluye variables demográficas de los anotadores como género y edad, lo cual aporta una dimensión sociocultural al análisis y permite estudiar potenciales sesgos en la percepción del contenido misógino (Plaza et al., 2024).

Los atributos del dataset EXIST 2024 se describen a continuación:

- `id_EXIST`: Un identificador único del tweet dentro del dataset EXIST.
- `tweet`: El texto completo del tweet en su formato original.
- `number_annotators`: El número de personas que han anotado el tweet. Este valor permite calcular el nivel de acuerdo entre anotadores.
- `annotators`: Un identificador único para cada uno de los anotadores que participaron en la clasificación del tweet.
- `gender_annotators`: El género de los diferentes anotadores. Los valores posibles son “F” (femenino) y “M” (masculino), permitiendo analizar si el género del anotador influye en la percepción del contenido sexista.

- `age_annotators`: El grupo de edad de los diferentes anotadores. Los valores posibles son: 18-22, 23-45 y 46+, permitiendo explorar diferencias generacionales en la percepción del sexismo.
- `labels_task1`: Un conjunto de etiquetas (una para cada anotador) que indican si el tuit contiene expresiones sexistas o se refiere a comportamientos sexistas. Los valores posibles son “YES” y “NO”.
- `labels_task2`: Un conjunto de etiquetas (una para cada anotador) que registra la intención de quien escribió el tuit. Las etiquetas posibles son: “DIRECT” (directo), “REPORTED” (reportado), “JUDGEMENTAL” (juicio), “-” (no aplica), y “UNKNOWN” (desconocido).
- `labels_task3`: Un conjunto de matrices de etiquetas (una para cada anotador) que indican el tipo o tipos de sexismo presentes en el tuit. Las etiquetas posibles son: “IDEOLOGICAL-INEQUALITY” (desigualdad ideológica), “STEREOTYPING-DOMINANCE” (estereotipos y dominación), “OBJECTIFICATION” (objetivación), “SEXUAL-VIOLENCE” (violencia sexual), “MISOGYNY-NON-SEXUAL-VIOLENCE” (misoginia y violencia no sexual), “-” (no aplica), y “UNKNOWN” (desconocido).
- `split`: Subconjunto dentro del conjunto de datos al que pertenece el tuit. Los valores posibles son “TRAIN” (entrenamiento), “DEV” (validación), “TEST” (prueba), seguidos de “EN” (inglés) o “ES” (español) para indicar el idioma.

3.2.3. Integración de los datasets

Ambos datasets fueron integrados y procesados para crear un corpus unificado que aprovecha las fortalezas de cada fuente. El dataset MEX-A3T proporciona categorías específicas de misoginia, mientras que EXIST aporta mayor diversidad de anotadores y variables sociodemográficas. Para mantener la consistencia, se conservó únicamente la información relevante para el análisis textual y las etiquetas de clasificación binaria (misógino/no misógino), dado que el objetivo principal de este trabajo es la clasificación binaria en lugar de la clasificación multiclase por categorías específicas.

Esta integración resultó en un corpus combinado de aproximadamente 6,903 tweets, proporcionando un volumen de datos suficiente para entrenar modelos robustos de aprendizaje profundo y permitiendo una validación más rigurosa de los resultados.

Si bien ambos datasets presentan distribuciones de clases aceptables de forma individual, el proceso de armonización de etiquetas descrito en esta sección resultó en que una proporción significativa de los registros no misóginos del dataset EXIST 2024 no pudiera ser incorporada al corpus fusionado, al no corresponderse con el esquema de etiquetado de MEX-A3T. Como consecuencia, el corpus combinado final presenta un desbalance de clases de 76.1 % misógino frente a 23.9 % no misógino, cuya atención mediante pesos de clase proporcionales se describe en el Capítulo 4.

3.3. Preprocesamiento de texto

Se realizaron diversas acciones de limpieza y preparación del texto para garantizar consistencia y eliminar ruido que pudiera afectar negativamente el desempeño de los modelos de clasificación. El preprocesamiento de texto en tareas de clasificación de redes sociales requiere un balance cuidadoso entre la eliminación de elementos irrelevantes y la preservación de características lingüísticas significativas para la detección de misoginia.

3.3.1. Eliminación de registros duplicados

Los registros duplicados pueden introducir sesgo en el entrenamiento al sobrerrepresentar ciertos patrones lingüísticos, además de inflar artificialmente las métricas de evaluación. Se identificaron y eliminaron tweets repetidos utilizando una comparación exacta del contenido del texto. Esta verificación se realizó aplicando funciones de detección de duplicados sobre la columna que contiene el mensaje.

El siguiente código Python implementa la detección y eliminación de duplicados:

```
df_clean = df.drop_duplicates(subset=[text_column])
final_count = len(df_clean)
print(f"Registros después de eliminar duplicados: {final_count}")
```

```
print(f"Registros eliminados: {initial_count - final_count}")
```

Los resultados del proceso de eliminación de duplicados para ambos datasets se presentan en las Figuras 3.1 y 3.2.

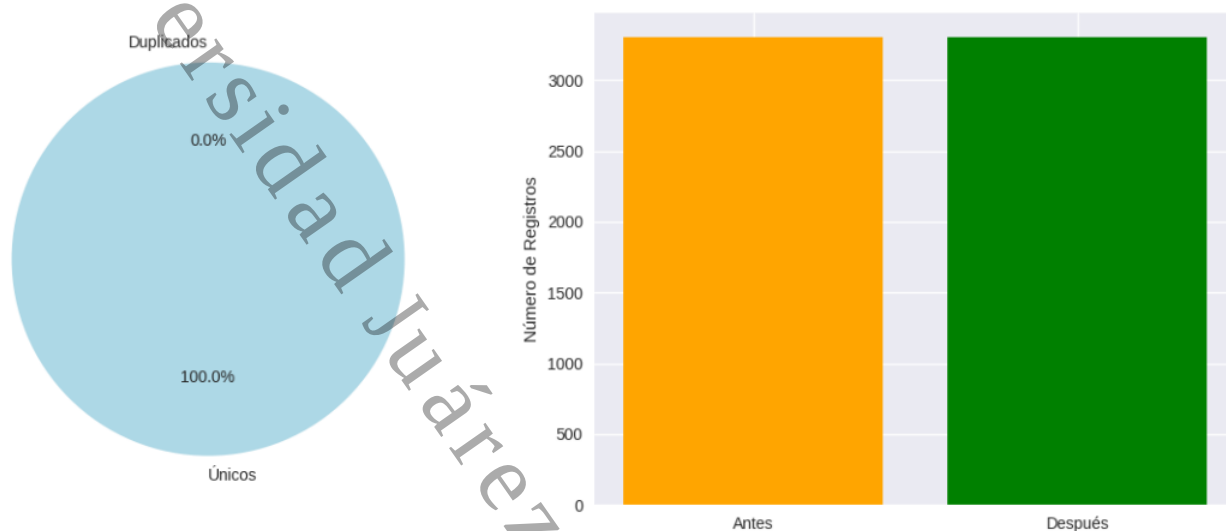


Figura 3.1. Análisis de registros duplicados en el dataset MEX-A3T

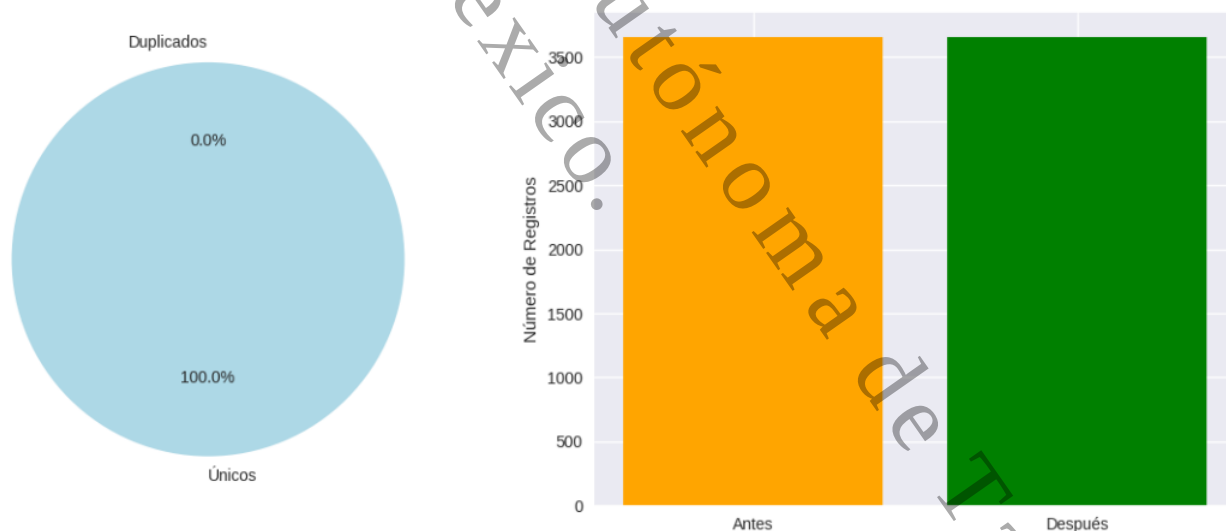


Figura 3.2. Análisis de registros duplicados en el dataset EXIST 2024.

La eliminación de duplicados garantiza que el modelo no aprenda patrones redundantes que sesguen su rendimiento hacia ciertos mensajes específicos en lugar de generalizar adecuadamente.

3.3.2. Eliminación de registros incompletos

Se revisó la presencia de valores nulos o vacíos, tanto en los textos como en las etiquetas. Los registros que no contienen texto o carecen de anotaciones deben ser eliminados, ya que no aportan información útil al proceso de entrenamiento y pueden causar errores durante el procesamiento.

El siguiente código implementa la detección de valores faltantes:

```
print("Valores■faltantes■por■columna:")
missing_info = df.isnull().sum()
print(missing_info)
empty_text = df[text_column].isna() | (df[text_column].str.strip() == '')
empty_count = empty_text.sum()
print(f"\nTextos■vacíos■o■nulos:■{empty_count}")
print(f"Porcentaje■de■textos■vacíos:■{(empty_count/initial_count)*100:.2f}%")
df_clean = df.dropna(subset=[text_column])
df_clean = df_clean[df_clean[text_column].str.strip() != '']
final_count = len(df_clean)
print(f"Registros■despues■de■limpieza:■{final_count}")
print(f"Registros■eliminados:■{initial_count-■final_count}")
```

Los resultados del análisis de registros incompletos se presentan en las Figuras 3.3 y 3.4. Afortunadamente, no se encontraron valores nulos en ninguno de los dos datasets, lo que evidencia la calidad del proceso de recopilación y anotación original.

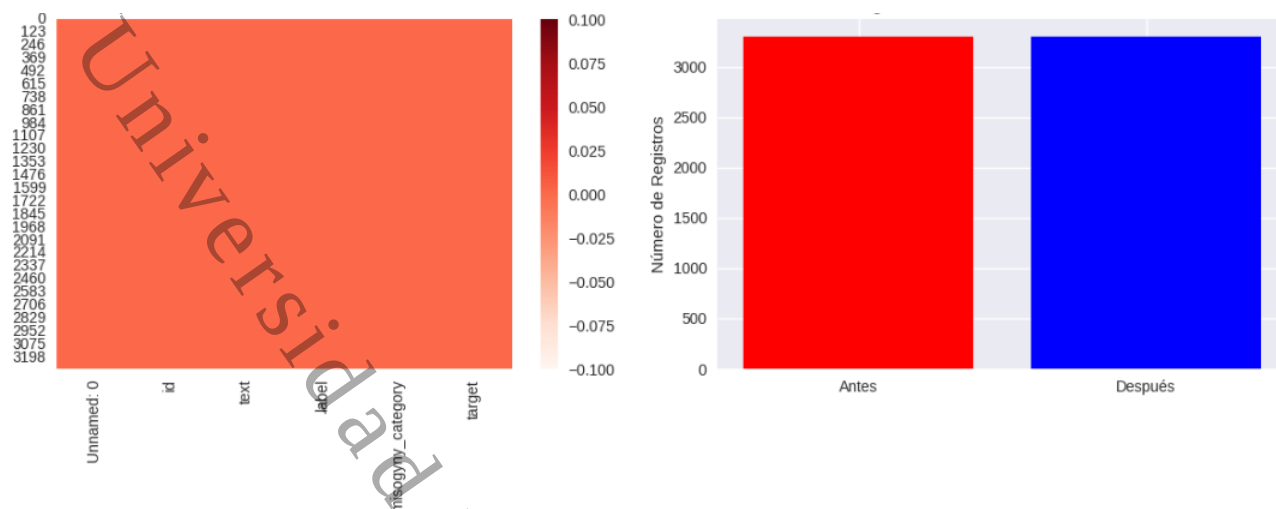


Figura 3.3. Verificación de registros incompletos en el dataset MEX-A3T.

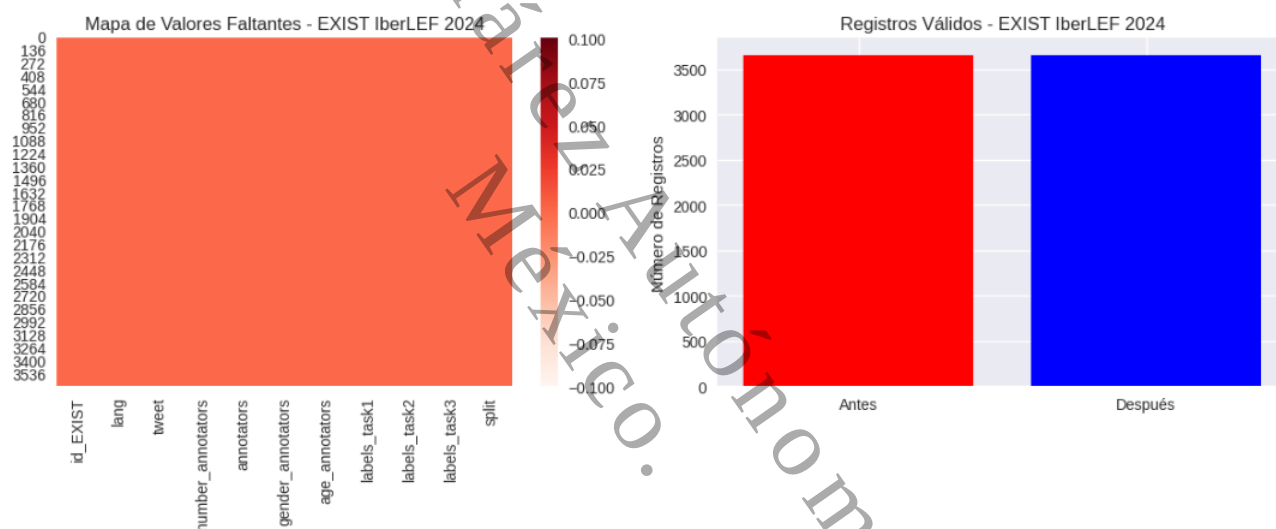


Figura 3.4. Verificación de registros incompletos en el dataset EXIST 2024.

3.3.3. Conversión a minúsculas

Para evitar que el modelo distinga entre formas mayúsculas y minúsculas de una misma palabra (por ejemplo, “Mujer” y “mujer”), se convirtió todo el contenido textual a minúsculas. Esta transformación estandariza el corpus y reduce significativamente la dimensionalidad del vocabulario, permitiendo que el modelo generalice mejor sobre el contenido semántico en lugar de características superficiales de formato.

Los resultados de la conversión a minúsculas se ilustran en las Figuras 3.5 y 3.6.

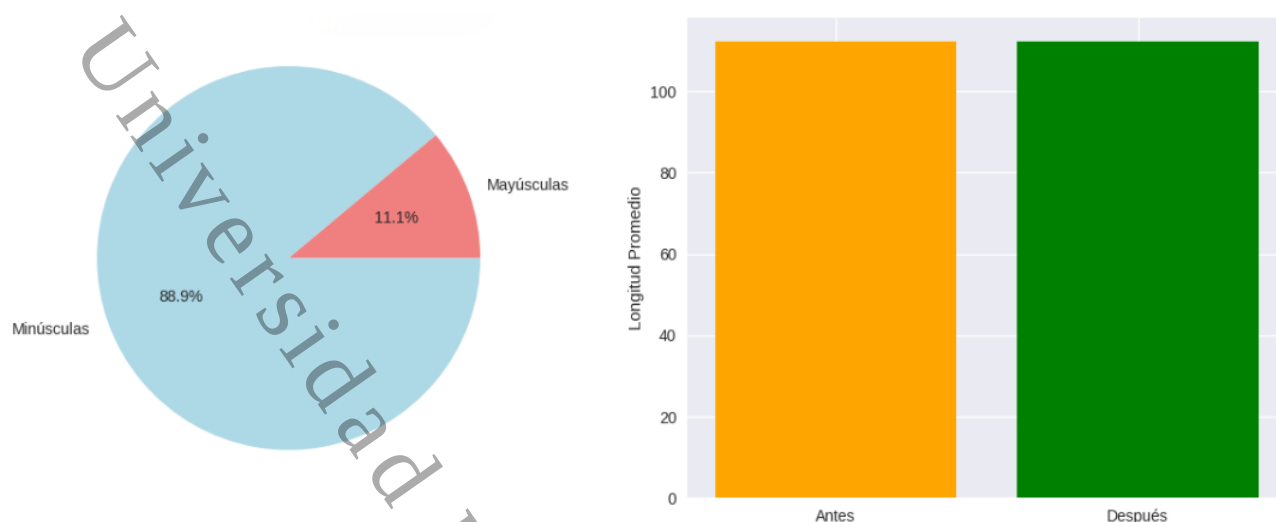


Figura 3.5. Estadísticas de conversión a minúsculas en el dataset MEX-A3T.

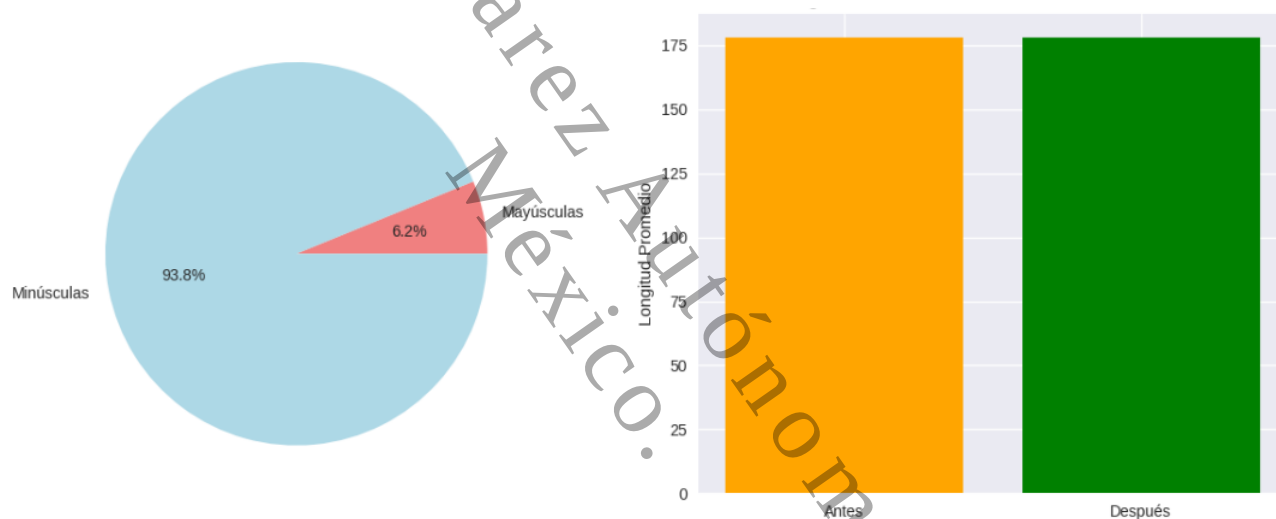


Figura 3.6. Estadísticas de conversión a minúsculas en el dataset EXIST 2024.

3.3.4. Eliminación de URLs

Los mensajes de redes sociales suelen contener enlaces a sitios web. Estas URLs no aportan valor semántico para la clasificación de misoginia en el contexto de este estudio y pueden introducir ruido en el modelo, por lo que fueron eliminadas utilizando expresiones regulares. La decisión de eliminar URLs se fundamenta en que el contenido misógino reside en el texto del mensaje mismo, no en los enlaces compartidos.

```
def analyze_url_removal(df, text_column, dataset_name):
```

```
    """ Eliminación de URLs """
```

```

print(f"\n===ELIMINACION DE URLs==={dataset_name}===")
texts = df[text_column].astype(str)

# Patron para URLs
url_pattern = r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-@.&+]|[*\(\)\,]|(?:%[0-9a-fA

# Encontrar URLs
urls_found = texts.str.findall(url_pattern)
total_urls = sum(len(url_list) for url_list in urls_found)
texts_with_urls = sum(1 for url_list in urls_found if len(url_list) > 0)

print(f"Total de URLs encontradas: {total_urls}")
print(f"Textos con URLs: {texts_with_urls}")
print(f"Porcentaje de textos con URLs: {(texts_with_urls / len(df)) * 100:.2f}%")

# Eliminar URLs
df_clean = df.copy()
df_clean[text_column] = texts.str.replace(url_pattern, '', regex=True)

return df_clean

```

Las Figuras 3.7 y 3.8 muestran las estadísticas de URLs encontradas y eliminadas en ambos datasets.

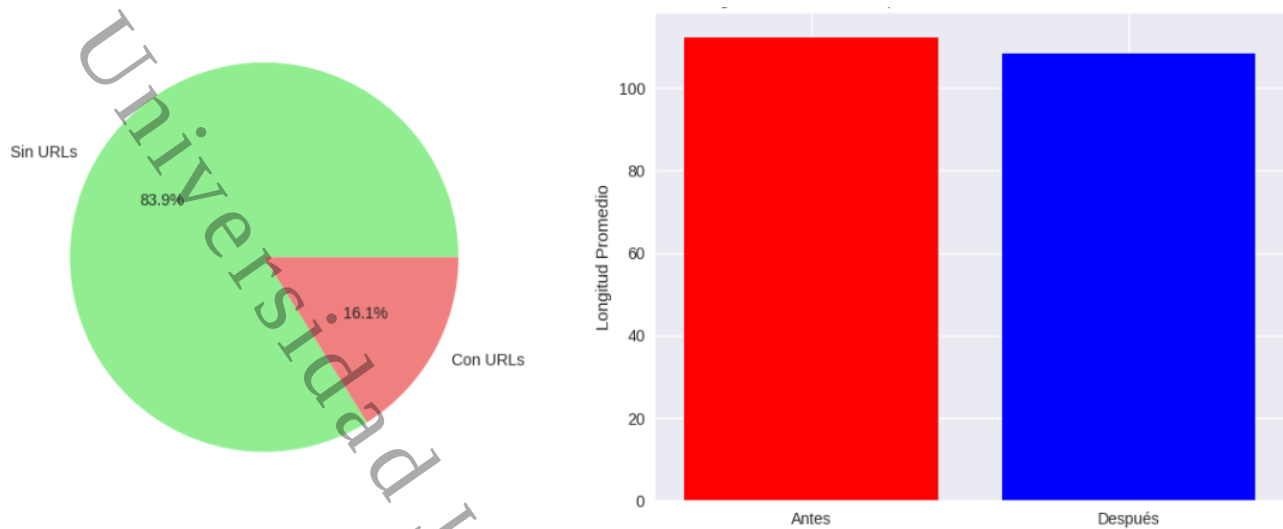


Figura 3.7. Análisis y eliminación de URLs en el dataset MEX-A3T.

3.3.5. Eliminación de menciones

Las menciones a otros usuarios (por ejemplo, @usuario) tampoco ofrecen información significativa para el análisis del contenido misógino per se, ya que la misoginia reside en el mensaje textual y no en las cuentas mencionadas. Por lo tanto, fueron removidas del texto utilizando expresiones regulares.

```
def analyze_mention_removal(df, text_column, dataset_name):
    """ Eliminación de menciones """
    print(f"\n=== ELIMINACION DE MENCIONES === {dataset_name} ===")

    texts = df[text_column].astype(str)

    # Patron para menciones (@usuario)
    mention_pattern = r'@[A-Za-z0-9_]+'

    # Encontrar menciones
    mentions_found = texts.str.findall(mention_pattern)
    total_mentions = sum(len(mention_list) for mention_list in mentions_found)
    texts_with_mentions = sum(1 for mention_list in mentions_found
```

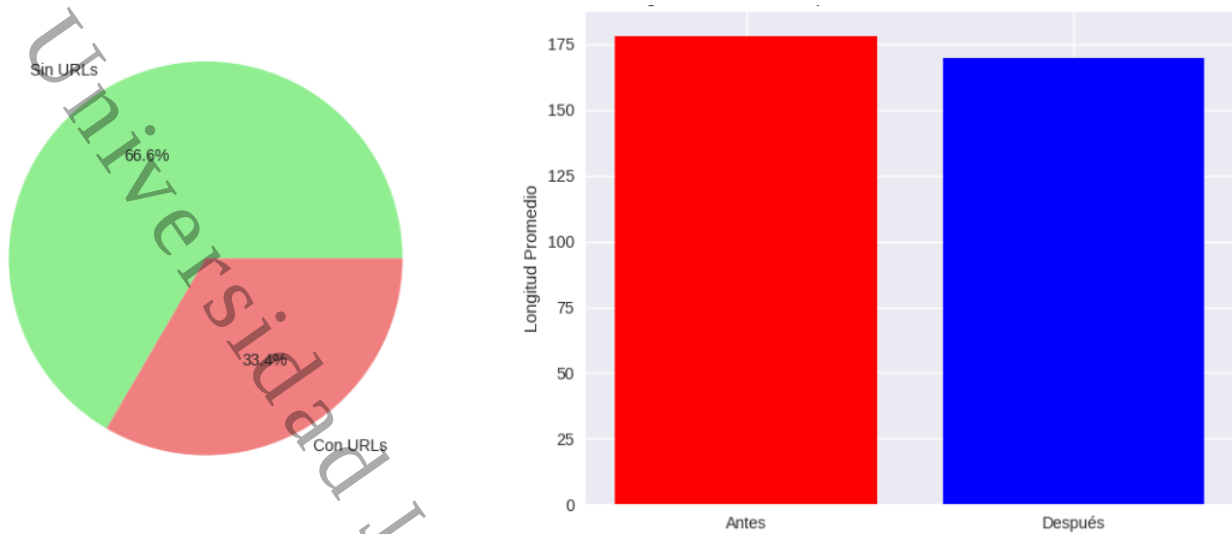


Figura 3.8. Análisis y eliminación de URLs en el dataset EXIST 2024.

```

if len(mention_list) > 0)

print(f"Total de menciones encontradas: {total_mentions}")
print(f"Textos con menciones: {texts_with_mentions}")
print(f"Porcentaje de textos con menciones: {(texts_with_mentions / len(df)) * 100}")

# Eliminar menciones
df_clean = df.copy()
df_clean[text_column] = texts.str.replace(mention_pattern, '', regex=True)

return df_clean

```

Los resultados de la eliminación de menciones se presentan en las Figuras 3.9 y 3.10.

3.3.6. Eliminación de caracteres especiales

Se eliminaron símbolos como signos de puntuación, emojis, hashtags y otros caracteres no alfabéticos, dejando únicamente texto limpio que pueda ser procesado de forma eficiente por los modelos de lenguaje. También se eliminaron múltiples espacios en blanco que pudieran haber quedado tras las eliminaciones previas. Esta decisión se tomó considerando que, aunque

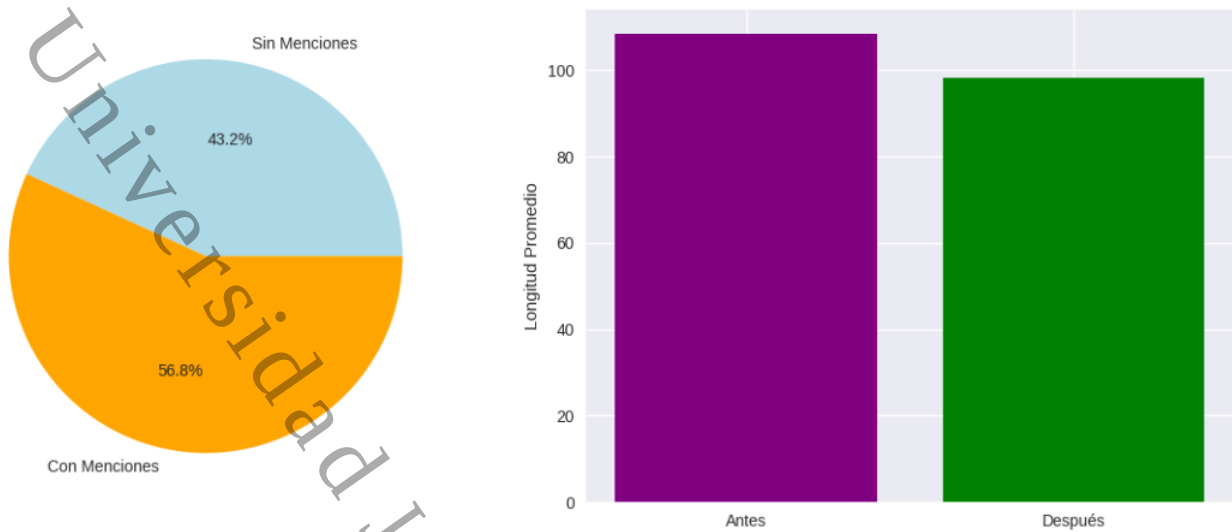


Figura 3.9. Análisis y eliminación de menciones en el dataset MEX-A3T.

los emojis y hashtags pueden contener información semántica, la arquitectura basada en BETO procesa principalmente texto alfabético estándar.

```
def analyze_special_chars_removal(df, text_column, dataset_name):
    """Eliminación de caracteres especiales"""
    print(f"\n=== ELIMINACION DE CARACTERES ESPECIALES - {dataset_name} ===")
    texts = df[text_column].astype(str)

    special_chars = texts.str.count(r'[\W\s]').sum()
    emojis = texts.str.count(r'[\W\s\u0000-\u007F]').sum()
    hashtags = texts.str.count(r'#\w+').sum()
    print(f"Caracteres especiales encontrados: {special_chars}")
    print(f"Caracteres no ASCII (emojis, etc.): {emojis}")
    print(f"Hashtags encontrados: {hashtags}")
    df_clean = df.copy()

    df_clean[text_column] = texts.str.replace(r'#\w+', '', regex=True)
    df_clean[text_column] = df_clean[text_column].str.replace(
        r'[\W\s]', ' ', regex=True)
```

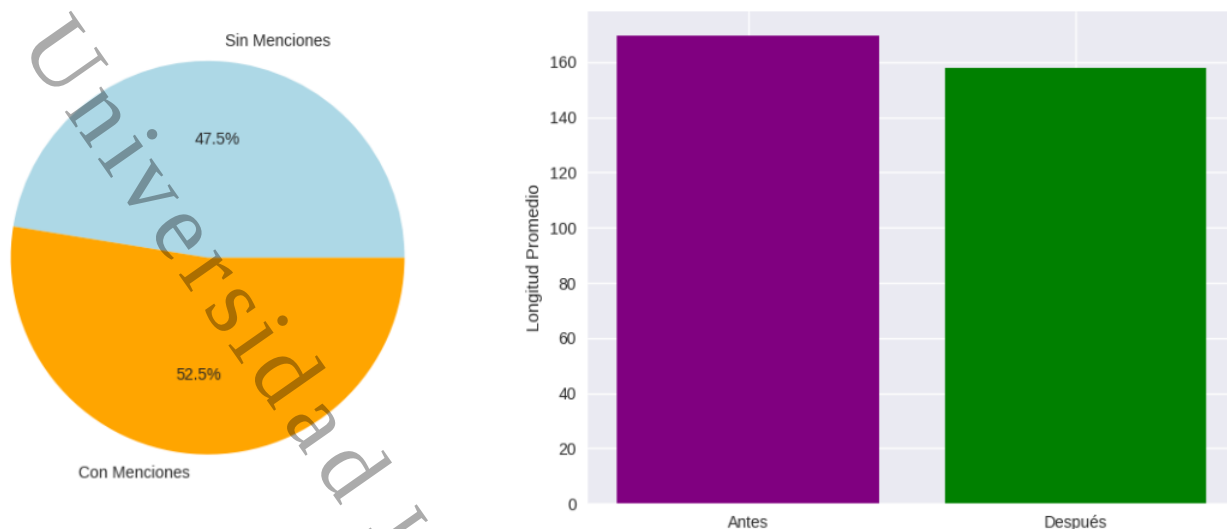


Figura 3.10. Análisis y eliminación de menciones en el dataset EXIST 2024.

```
df_clean[text_column] = df_clean[text_column].str.replace(
    r'\s+', '█', regex=True)
df_clean[text_column] = df_clean[text_column].str.strip()

return df_clean
```

Las Figuras 3.11 y 3.12 muestran las estadísticas de caracteres especiales procesados.

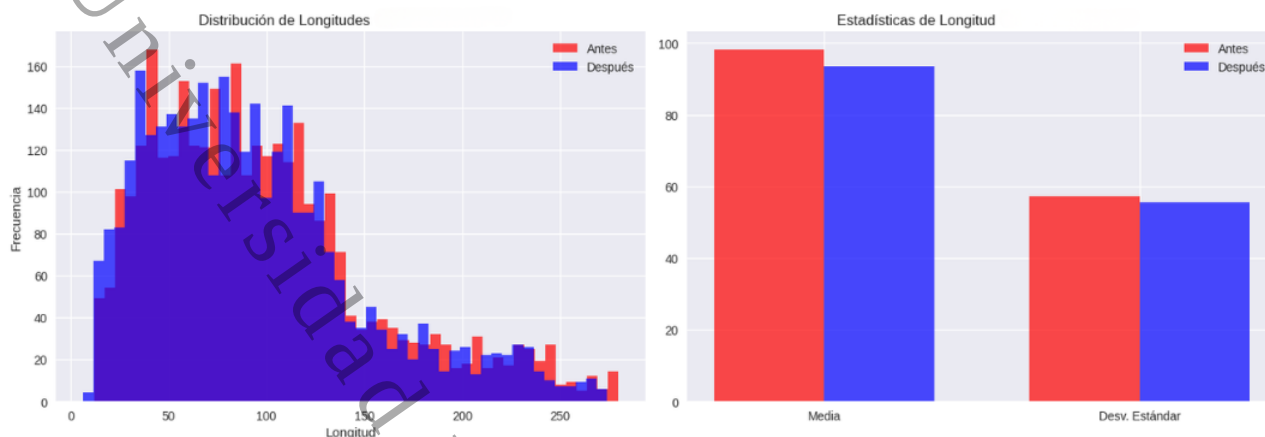


Figura 3.11. Análisis y eliminación de caracteres especiales en el dataset MEX-A3T.

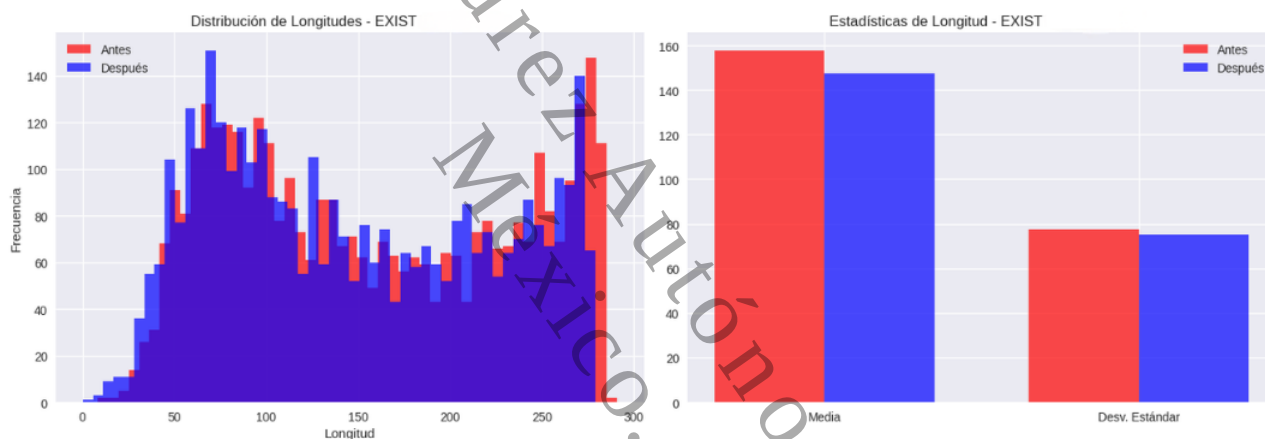


Figura 3.12. Análisis y eliminación de caracteres especiales en el dataset EXIST 2024.

3.4. Análisis estadístico

Una vez completado el proceso de limpieza y preprocesamiento de los datos, se procedió a realizar un análisis estadístico descriptivo de ambos conjuntos de datos para comprender mejor las características del corpus y validar la calidad de la preparación realizada. Este análisis proporciona información crucial sobre la distribución de clases, el balance del dataset y otros aspectos que pueden influir en el rendimiento del modelo.

3.4.1. Distribución de clases

Se analizó la distribución de las etiquetas de clasificación en ambos datasets para identificar posibles desbalances que pudieran afectar el entrenamiento del modelo. El desbalance de clases es un problema común en tareas de detección de discurso de odio, donde típicamente la clase mayoritaria (contenido no ofensivo) supera significativamente a la clase minoritaria (contenido ofensivo), lo que puede sesgar el modelo hacia la predicción de la clase mayoritaria.

La Figura 3.13 muestra la distribución de clases en el dataset MEX-A3T.

Los resultados mostraron que en el dataset MEX-A3T, el 52.3 % de los tweets corresponden a contenido no misógino (clase 0) mientras que el 47.7 % son clasificados como misóginos (clase 1). Esta distribución relativamente balanceada es favorable para el entrenamiento del modelo, ya que reduce el riesgo de sesgo hacia la clase mayoritaria y permite que el modelo aprenda patrones discriminativos de ambas clases de manera equitativa. El ratio aproximado de 1.1:1 se considera óptimo para entrenamiento sin necesidad de técnicas adicionales de balanceo como oversampling o class weighting.

La Figura 3.14 presenta la distribución de clases en el dataset EXIST 2024.

En el caso del dataset EXIST 2024, se observó una distribución del 59.7 % para contenido no sexista y 40.3 % para contenido sexista, manteniendo también un balance adecuado para el aprendizaje automático. Con un ratio de aproximadamente 1.4:1, este dataset presenta un desbalance ligeramente mayor que MEX-A3T pero aún dentro de rangos manejables que no requieren técnicas especiales de balanceo.

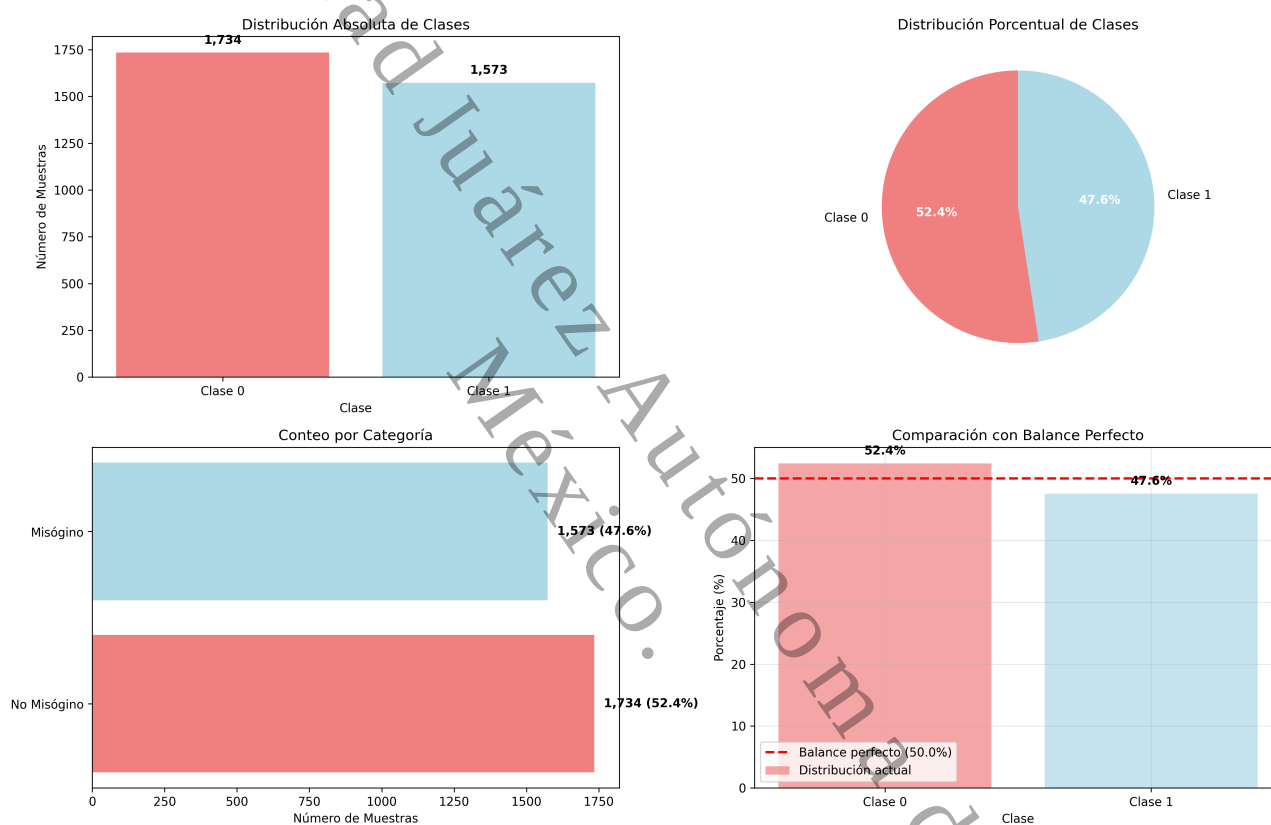


Figura 3.13. Distribución de clases en el dataset MEX-A3T. Se observa una distribución relativamente balanceada con 52.3% de mensajes no misóginos y 47.7% de mensajes misóginos.

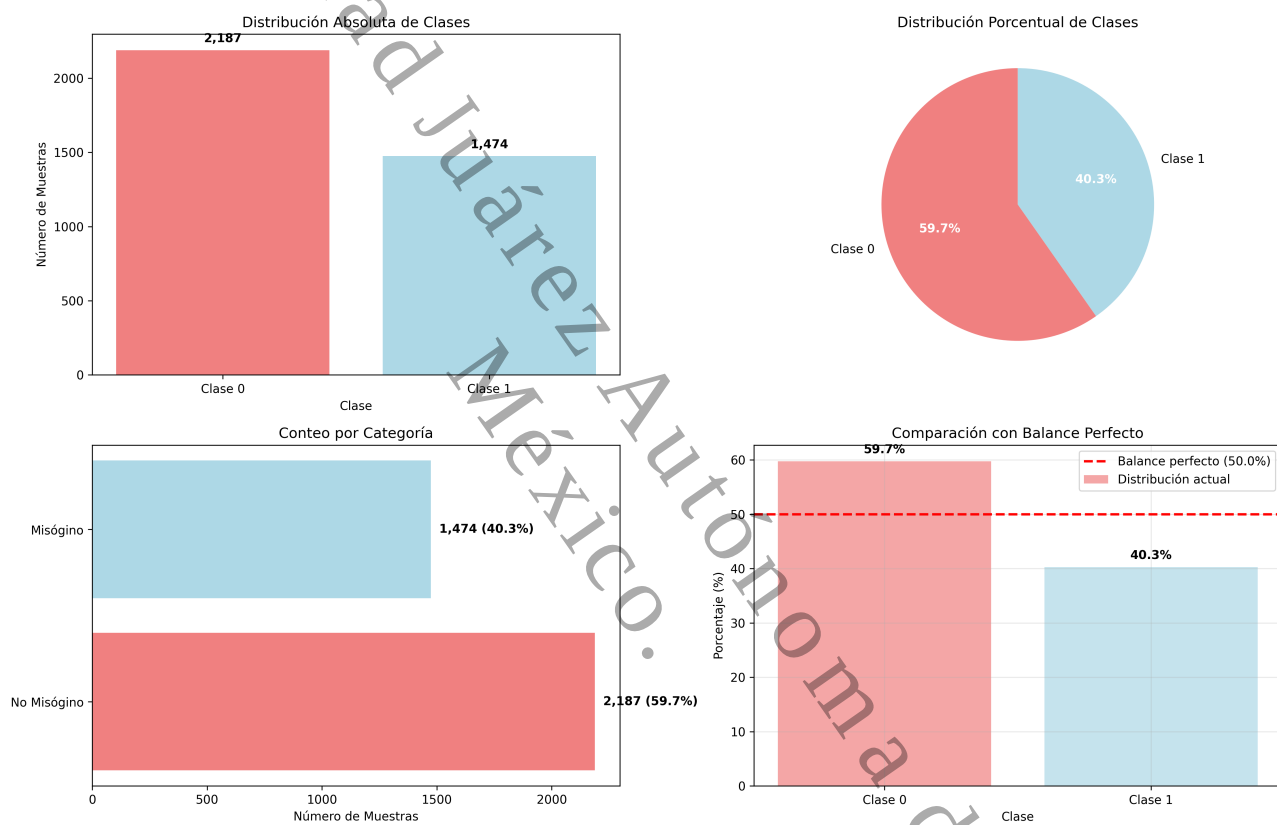


Figura 3.14. Distribución de clases en el dataset EXIST 2024. La distribución muestra 59.7% de contenido no sexista y 40.3% de contenido sexista.

Capítulo 4

Experimentos y Resultados

4.1. Diseño experimental

El diseño experimental de este estudio se estructura a partir de los conjuntos de datos descritos en el Capítulo 3, los cuales fueron sometidos a un riguroso proceso de limpieza y preprocesamiento. En esta etapa, el enfoque se centra en la configuración de los modelos neuronales unimodales, la definición de las métricas de evaluación y la validación de los resultados.

Los experimentos se organizaron en dos fases. En la **Fase 1** se entrenaron y evaluaron las tres arquitecturas (LSTM, CNN y BETO) utilizando cada dataset por separado, con el objetivo de analizar el comportamiento individual de cada modelo y establecer una línea base de rendimiento. En la **Fase 2** se incorporó XLM-RoBERTa large y se evaluaron las cuatro arquitecturas sobre el dataset fusionado resultante de la integración de MEX-A3T y EXIST 2024, con un total de 6,903 mensajes tras la eliminación de duplicados.

La distribución real del corpus fusionado presenta un desbalance de clases de 76.1 % mensajes misóginos frente a 23.9 % no misóginos, con un ratio de desbalance de aproximadamente 1:3.18. Este desbalance fue identificado durante la fase experimental y atendido mediante la aplicación de pesos de clase proporcionales (*class weights*) calculados con el método `balanced` de `scikit-learn`.

Se consideraron las siguientes etapas metodológicas:

División de datos: Los conjuntos fueron separados en particiones de entrenamiento (70 %), va-

lidación (15 %) y prueba (15 %), garantizando representación proporcional de ambas clases mediante estratificación.

Modelado: En la Fase 1 se evaluaron LSTM, CNN y BETO. En la Fase 2 se incorporó XLM-RoBERTa large para determinar si su mayor capacidad paramétrica compensaba la falta de especialización monolingüe en español.

Control del sobreajuste: Durante la fase exploratoria se detectó sobreajuste en las arquitecturas evaluadas, manifestado como divergencia entre las curvas de pérdida de entrenamiento y validación. En respuesta, se eliminó la regularización L2 agresiva que, en combinación con *class weights* y datasets de tamaño moderado, causaba inestabilidad en el gradiente. El control del sobreajuste se delegó exclusivamente a mecanismos de *dropout*, *Batch Normalization* y *early stopping*.

Evaluación: Se aplicaron métricas estándar de clasificación binaria (accuracy, precision, recall, F1-score y Macro-F1), complementadas con matrices de confusión para analizar patrones de error. Dado el desbalance del corpus fusionado (76/24), el Macro-F1 se adopta como métrica principal de comparación entre arquitecturas.

4.2. Configuración de modelos

4.2.1. Modelos secuenciales

LSTM (Long Short-Term Memory): Se implementó una red neuronal recurrente LSTM con el objetivo de capturar dependencias secuenciales en los mensajes. En la Fase 1 la arquitectura incluyó embeddings de dimensión 300, 128 unidades LSTM, una capa densa de 64 unidades con activación ReLU, dropout de 0.5 y salida sigmoide. El optimizador empleado fue Adam con tasa de aprendizaje de 1×10^{-3} , batch size de 64 y early stopping con paciencia de 3 épocas sobre un máximo de 20.

Para la Fase 2 se modificó la arquitectura para atender el desbalance del corpus fusionado: embeddings de dimensión 128, SpatialDropout1D con tasa 0.1, 64 unidades LSTM con dropout interno de 0.3 (recurrente 0.2), *Batch Normalization*, capa densa de 32 unidades con ReLU, dro-

pout de 0.3 y salida sigmoide. Se eliminó la regularización L2 de todas las capas.

CNN para texto: En la Fase 1 se implementó una CNN con embeddings de dimensión 300, tres ramas convolucionales con filtros de tamaños 3, 4 y 5 (100 filtros cada una), dropout de 0.5, optimizador Adam con tasa 1×10^{-3} y batch size de 64.

Para la Fase 2 se ajustó la arquitectura basada en kim2014cnn: embeddings de dimensión 128, SpatialDropout1D con tasa 0.2, tres ramas convolucionales con filtros de tamaños 2, 3 y 4 (64 filtros cada una) seguidas de GlobalMaxPooling1D, *BatchNormalization*, capa densa de 64 unidades con ReLU, dropout de 0.4 y salida sigmoide. Se eliminó la regularización L2 de todas las capas.

4.2.2. Modelos basados en Transformers

BETO: En ambas fases se utilizó la versión dccuchile/bert-base-spanish-wwm-cased con 110 millones de parámetros (Cañete et al., 2020). En la Fase 1: tasa de aprendizaje 2×10^{-5} , batch size 32, 4 épocas, optimizador AdamW con weight decay 0.01, warmup lineal del 10% y dropout por defecto (0.1). En la Fase 2: tasa de aprendizaje 1×10^{-5} , batch size 16, máximo 3 épocas con early stopping de paciencia 2, AdamW con weight decay 0.05, warmup del 20%, dropout 0.2, y función de pérdida CrossEntropyLoss con pesos de clase proporcionales (Clase 0: 2.0872, Clase 1: 0.6575).

XLm-RoBERTa large (solo Fase 2): modelo multilingüe con 560 millones de parámetros entrenado en 100 idiomas (Conneau et al., 2020). Configuración idéntica a BETO en la Fase 2.

4.3. Entorno de ejecución

Los experimentos se llevaron a cabo en Google Colaboratory con acceso a GPU NVIDIA Tesla T4 (16 GB de VRAM) y Python 3.12. Las bibliotecas utilizadas fueron:

- Transformers 4.x (HuggingFace) — BETO y XLm-RoBERTa large.
- PyTorch 2.x — clasificadores basados en transformers.
- TensorFlow 2.19 y Keras — redes LSTM y CNN.

- scikit-learn 1.x — métricas, matrices de confusión y pesos de clase.
- Pandas y NumPy — manipulación y análisis de datos.
- Matplotlib y Seaborn — visualización de resultados.

4.4. Métricas de evaluación

Para evaluar el desempeño de los modelos se seleccionaron métricas ampliamente utilizadas en tareas de clasificación binaria y recomendadas en competencias como IberEval e IberLEF. El Macro-F1 se adopta como métrica principal de comparación entre arquitecturas, ya que promedia el F1-score de cada clase de forma independiente sin ponderar por frecuencia, siendo insensible al desbalance. A efectos de la verificación de la hipótesis planteada en el Capítulo 1, el término *precisión* se operacionaliza mediante el Macro-F1, dado que esta métrica refleja de manera más fidedigna el rendimiento real del clasificador ante corpus desbalanceados.

Las métricas empleadas fueron:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

Precision:

$$Precision = \frac{TP}{TP + FP} \quad (4.2)$$

Recall:

$$Recall = \frac{TP}{TP + FN} \quad (4.3)$$

F1-score:

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4.4)$$

Macro-F1: Promedio del F1-score calculado de manera independiente para cada clase; métrica principal de este trabajo.

Matriz de confusión: Representación tabular de TP, FP, TN y FN que permite identificar los tipos de errores más comunes.

4.5. Resultados experimentales

4.5.1. Observación sobre sobreajuste en experimentos exploratorios

Durante la fase exploratoria del diseño experimental se detectó sobreajuste en las arquitecturas evaluadas. En los modelos CNN y BETO, las curvas de pérdida de entrenamiento descendían hacia valores cercanos a cero mientras que la pérdida de validación aumentaba a partir de las épocas intermedias. En el modelo LSTM, el problema fue más severo en el corpus fusionado: la combinación de regularización L2 agresiva con pesos de clase mal calibrados produjo colapso del gradiente, haciendo que el modelo convergiera trivialmente hacia la clase mayoritaria con Recall de 0 % para la clase no misógina.

En respuesta se realizó una revisión sistemática de las configuraciones. La estrategia adoptada fue eliminar la regularización L2 y delegar el control del sobreajuste exclusivamente a dropout, BatchNormalization y early stopping. Los resultados reportados en las secciones siguientes corresponden a las configuraciones corregidas.

4.5.2. Fase 1: Resultados con datasets individuales

En esta primera fase experimental, se entrenaron y evaluaron las tres arquitecturas (LSTM, CNN y BETO) utilizando cada dataset por separado. Este enfoque permite analizar el comportamiento de cada modelo ante las características específicas de cada fuente de datos y comprender cómo la cantidad y diversidad de los datos afectan el rendimiento.

4.5.2.1. Resultados en el dataset MEX-A3T

El dataset MEX-A3T, compuesto por 3,307 tweets en español mexicano, fue utilizado como primer conjunto de evaluación. Los datos fueron divididos en: 70 % para entrenamiento (2,315 tweets), 15 % para validación (496 tweets) y 15 % para prueba (496 tweets), manteniendo la proporción de clases (52.3 % no misógino, 47.7 % misógino) mediante estratificación.

Modelo BETO: Se utilizó la arquitectura `dccuchile/bert-base-spanish-wwm-cased` con la configuración de Fase 1 descrita en la sección 4.2. El modelo fue entrenado en GPU NVIDIA Tesla

T4 con un tiempo aproximado de 3 horas 15 minutos. Los resultados se presentan en la Tabla 4.1.

Tabla 4.1. Resultados del modelo BETO entrenado con el dataset MEX-A3T.

Métrica	Valor
Accuracy	82.09 %
Precision	81.93 %
Recall	82.26 %
F1-score	82.09 %
Macro-F1	82.09 %

El modelo alcanzó un Macro-F1 de 82.09 %, demostrando un desempeño competitivo considerando el tamaño relativamente limitado del dataset de entrenamiento.

Modelo CNN: La arquitectura CNN se entrenó con la configuración de Fase 1. El tiempo de entrenamiento fue aproximadamente 8 minutos, evidenciando la mayor eficiencia computacional de las arquitecturas convolucionales. Los resultados se presentan en la Tabla 4.2.

Tabla 4.2. Resultados del modelo CNN entrenado con el dataset MEX-A3T.

Métrica	Valor
Accuracy	78.43 %
Precision	79.21 %
Recall	76.89 %
F1-score	78.03 %
Macro-F1	78.43 %

CNN alcanzó un Macro-F1 de 78.43 %, inferior a BETO pero con una eficiencia computacional significativamente mayor.

Modelo LSTM: El modelo LSTM se configuró con la arquitectura de Fase 1. Los resultados se presentan en la Tabla 4.3.

Tabla 4.3. Resultados del modelo LSTM entrenado con el dataset MEX-A3T.

Métrica	Valor
Accuracy	74.19 %
Precision	75.84 %
Recall	71.23 %
F1-score	73.46 %
Macro-F1	74.19 %

LSTM obtuvo el rendimiento más bajo con 74.19 % de Macro-F1, evidenciando las limitaciones de las redes recurrentes sin preentrenamiento con datasets de tamaño moderado.

4.5.2.2. Resultados en el dataset EXIST 2024

El dataset EXIST 2024, con 3,660 tweets, fue dividido en: 70 % entrenamiento (2,562 tweets), 15 % validación (549 tweets) y 15 % prueba (549 tweets), manteniendo la distribución de clases (59.7 % no sexista, 40.3 % sexista) mediante estratificación.

Modelo BETO: Se mantuvo la misma configuración de hiperparámetros de Fase 1. Los resultados se presentan en la Tabla 4.4.

Tabla 4.4. Resultados del modelo BETO entrenado con el dataset EXIST 2024.

Métrica	Valor
Accuracy	80.69 %
Precision	78.92 %
Recall	81.45 %
F1-score	80.16 %
Macro-F1	80.69 %

BETO alcanzó un Macro-F1 de 80.69 %, ligeramente inferior al obtenido con MEX-A3T (82.09 %). Esta diferencia puede atribuirse a la mayor complejidad y diversidad de anotaciones del dataset EXIST, que incluye múltiples dimensiones de clasificación anotadas por varios evaluadores.

Modelo CNN: Los resultados de CNN en el dataset EXIST se presentan en la Tabla 4.5.

Tabla 4.5. Resultados del modelo CNN entrenado con el dataset EXIST 2024.

Métrica	Valor
Accuracy	76.87 %
Precision	74.56 %
Recall	78.34 %
F1-score	76.40 %
Macro-F1	76.87 %

CNN obtuvo un Macro-F1 de 76.87 %, consistente con el resultado en MEX-A3T (78.43 %), demostrando estabilidad de rendimiento a pesar de las diferencias entre datasets.

Tabla 4.6. Resultados del modelo LSTM entrenado con el dataset EXIST 2024.

Métrica	Valor
Accuracy	72.31 %
Precision	70.45 %
Recall	73.89 %
F1-score	72.13 %
Macro-F1	72.31 %

Modelo LSTM: Los resultados de LSTM en EXIST se muestran en la Tabla 4.6.

LSTM obtuvo un Macro-F1 de 72.31 %, el más bajo entre las tres arquitecturas y ligeramente inferior al resultado en MEX-A3T (74.19 %), confirmando que las arquitecturas recurrentes sin preentrenamiento son las más sensibles a la reducción del tamaño del corpus.

4.5.2.3. Análisis comparativo de la Fase 1

La Tabla 4.7 resume los resultados de Macro-F1 obtenidos en la Fase 1 para facilitar la comparación entre arquitecturas y datasets.

Tabla 4.7. Comparación de Macro-F1 en la Fase 1 (datasets individuales).

Modelo	MEX-A3T	EXIST 2024	Mejor individual
BETO	82.09 %	80.69 %	82.09 %
CNN	78.43 %	76.87 %	78.43 %
LSTM	74.19 %	72.31 %	74.19 %

Los resultados de la Fase 1 revelan una jerarquía consistente entre arquitecturas: BETO supera a CNN en aproximadamente 3.5–4 puntos de Macro-F1, y CNN supera a LSTM en aproximadamente 4 puntos, independientemente del dataset utilizado. Esta consistencia sugiere que las diferencias de rendimiento son atribuibles a las arquitecturas y no a características específicas de los datos.

4.5.3. Fase 2: Resultados con dataset fusionado

4.5.3.1. Configuración experimental del dataset fusionado

El dataset fusionado resultante de la integración de MEX-A3T y EXIST 2024 contiene 6,903 mensajes con distribución 76.1 % misógino y 23.9 % no misógino. Para la integración de EXIST 2024, se utilizaron los registros correspondientes a la tarea de detección de clase minoritaria,

que por diseño contiene exclusivamente ejemplos de contenido misógino. Como consecuencia, los 1,653 ejemplos no misóginos del corpus fusionado provienen íntegramente de MEX-A3T. Las etiquetas fueron armonizadas de la siguiente manera:

- MEX-A3T: `misogynous = 1` → misógino; `misogynous = 0` → no misógino.
- EXIST 2024: `labels_task1 = 'YES'` → misógino.

El dataset se dividió estratificadamente en: 70 % entrenamiento (4,834 mensajes), 15 % validación (1,033 mensajes) y 15 % prueba (1,036 mensajes).

4.5.3.2. Modelo BETO

El modelo BETO fue entrenado mediante fine-tuning con la configuración de Fase 2, incorporando pesos de clase proporcionales al desbalance (Clase 0: 2.0872, Clase 1: 0.6575). El entrenamiento fue detenido en la época 2 por early stopping (`val_loss = 0.2678`). Los resultados se presentan en la Tabla 4.8.

Tabla 4.8. Resultados del modelo BETO entrenado con el dataset fusionado.

Métrica	Valor
Accuracy	88.42 %
Precision	95.13 %
Recall	89.21 %
F1-Score	92.08 %
Macro-F1	85.28 %
Recall No Misógino	85.48 %
Recall Misógino	89.21 %

BETO alcanzó un Macro-F1 de 85.28 %, superando el umbral del 85 % establecido en la hipótesis, y una mejora de 3.19 puntos sobre su mejor resultado individual (82.09 % en MEX-A3T). Con 29 falsos positivos y 85 falsos negativos sobre 1,036 mensajes de prueba, BETO demostró la mayor capacidad de generalización entre todas las arquitecturas evaluadas.

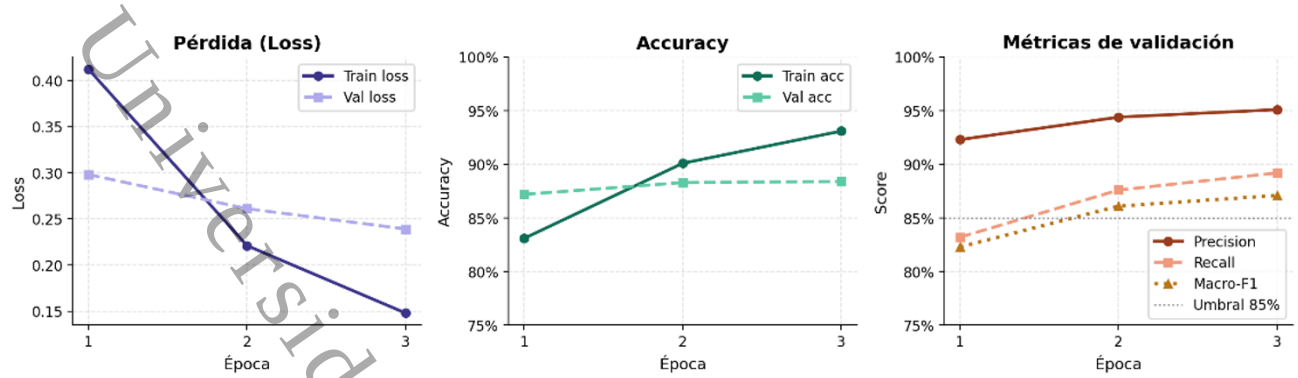


Figura 4.1. Curvas de entrenamiento del modelo BETO con el dataset fusionado. El modelo alcanza convergencia estable en 3 épocas con descenso consistente de la pérdida de entrenamiento hasta 0.15 y de validación hasta 0.24, sin divergencia entre ambas curvas. La accuracy de validación se mantiene estable en 88% desde la primera época. Las métricas de validación muestran que el Macro-F1 supera claramente el umbral del 85% en la época final, con Precision 95% y Recall 89%, confirmando la verificación de la hipótesis planteada. La línea punteada indica el umbral de 85% establecido en la hipótesis.

4.5.3.3. Modelo XLM-RoBERTa large

XLM-RoBERTa large fue evaluado únicamente con el dataset fusionado, con el objetivo de determinar si su mayor capacidad paramétrica (560 millones de parámetros) compensaba la falta de especialización monolingüe en español. Los resultados se presentan en la Tabla 4.9.

Tabla 4.9. Resultados del modelo XLM-RoBERTa large entrenado con el dataset fusionado.

Métrica	Valor
Accuracy	87.45%
Precision	95.57%
Recall	87.56%
F1-Score	91.39%
Macro-F1	84.13%
Recall No Misógino	87.10%
Recall Misógino	87.56%

XLM-RoBERTa large alcanzó un Macro-F1 de 84.13%, ligeramente inferior a BETO a pesar de contar con 5 veces más parámetros. Obtuvo la mayor Precision entre todos los modelos evaluados (95.57%), con solo 32 falsos positivos.

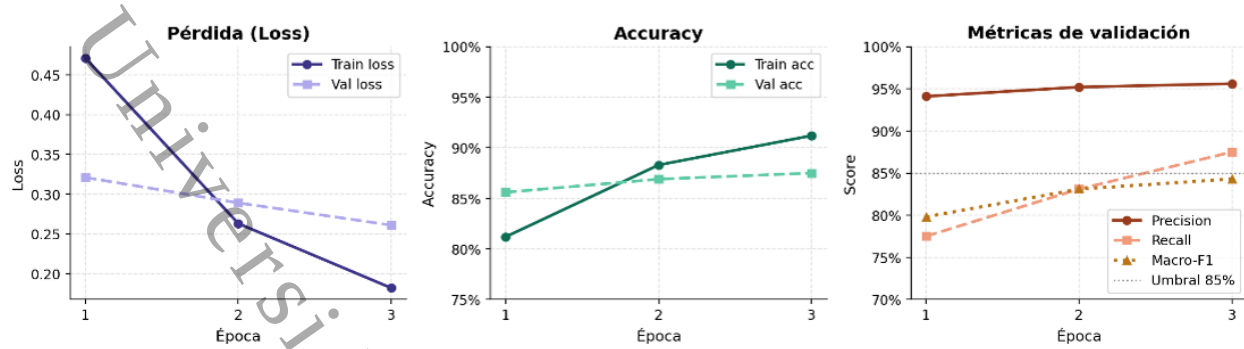


Figura 4.2. Curvas de entrenamiento del modelo XLM-RoBERTa large con el dataset fusionado. El modelo converge en tan solo 3 épocas con una reducción limpia y estable de la pérdida tanto en entrenamiento como en validación, sin señales de sobreajuste. La accuracy de validación se estabiliza en 87% y las métricas de validación muestran una Precision consistentemente alta (95%) a lo largo de todo el entrenamiento. El Macro-F1 muestra una tendencia creciente que se aproxima al umbral del 85% en la época final, sin llegar a superarlo de forma consolidada, lo que es consistente con el resultado reportado en la Tabla 4.9 (Macro-F1 = 84.13%).

4.5.3.4. Modelo CNN

La arquitectura CNN fue entrenada con filtros paralelos de tamaños 2, 3 y 4, sin regularización L2 y con pesos de clase proporcionales. Los resultados se presentan en la Tabla 4.10.

Tabla 4.10. Resultados del modelo CNN entrenado con el dataset fusionado.

Métrica	Valor
Accuracy	81.85 %
Precision	93.99 %
Recall	81.35 %
F1-Score	87.21 %
Macro-F1	77.99 %
Recall No Misógino	83.47 %
Recall Misógino	81.35 %

CNN alcanzó un Macro-F1 de 77.99%, inferior en 0.44 puntos respecto a su mejor resultado individual (78.43% en MEX-A3T). A diferencia de BETO, la fusión de datasets no representó una mejora para CNN, lo que sugiere que la arquitectura convolucional sin preentrenamiento no aprovecha la diversidad lingüística adicional del corpus combinado de la misma manera que los modelos transformer.

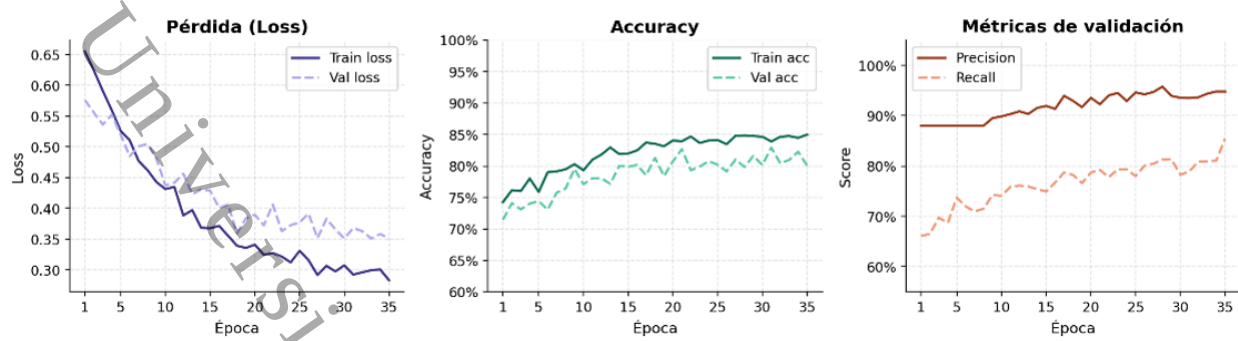


Figura 4.3. Curvas de entrenamiento del modelo CNN con el dataset fusionado. La gráfica de pérdida muestra una reducción consistente del error de entrenamiento a lo largo de las 35 épocas, mientras que la pérdida de validación desciende con mayor variabilidad, evidenciando un sobreajuste leve pero controlado por los mecanismos de *dropout* y *early stopping*. La accuracy de validación se estabiliza en torno al 82 %, mientras que las métricas de validación muestran una Precisión alta y estable (93 %) con un Recall que mejora progresivamente hasta aproximadamente 82 %, reflejando el efecto de los pesos de clase proporcionales sobre la detección de la clase minoritaria.

4.5.3.5. Modelo LSTM

La arquitectura LSTM fue evaluada con múltiples configuraciones en el corpus fusionado. En todas las configuraciones exploradas, el modelo presentó inestabilidad en la convergencia, colapsando hacia la predicción trivial de la clase mayoritaria (Recall No Misógino = 0 %, TN = 0). Este comportamiento se atribuye a la combinación del desbalance real del corpus (ratio 1:3.18) con la ausencia de preentrenamiento masivo en español, que limita la capacidad del modelo para construir representaciones semánticas discriminativas a partir del corpus de entrenamiento disponible. Dado que ninguna configuración explorada produjo convergencia estable, el modelo LSTM no es reportado como resultado experimental final en la Fase 2.

4.5.3.6. Análisis comparativo de la Fase 2

La Tabla 4.11 resume los resultados obtenidos por las arquitecturas que lograron convergencia estable en el dataset fusionado.

Tabla 4.11. Comparación de resultados entre arquitecturas con dataset fusionado.

Modelo	Accuracy	Macro-F1	FP	FN
BETO	88.42 %	85.28 %	29	85
XLM-RoBERTa large	87.45 %	84.13 %	32	98
CNN	81.85 %	77.99 %	41	147
LSTM	Sin convergencia estable			

BETO supera a XLM-RoBERTa large en Macro-F1 por 1.15 puntos porcentuales a pesar de

contar con 450 millones de parámetros menos. Ambos modelos transformer superan ampliamente a CNN, con diferencias de 7.29 y 6.14 puntos en Macro-F1 respectivamente.

4.5.4. Comparación detallada: BETO, XLM-RoBERTa large y CNN

La Tabla 4.12 presenta el perfil completo de los tres modelos que alcanzaron convergencia estable, incluyendo parámetros, tipo de preentrenamiento y verificación de la hipótesis.

Tabla 4.12. Comparación detallada entre los modelos con dataset fusionado.

Característica	BETO	XLM-RoBERTa	CNN
Accuracy (%)	88.42	87.45	81.85
Macro-F1 (%)	85.28	84.13	77.99
Precision (%)	95.13	95.57	93.99
Recall (%)	89.21	87.56	81.35
Recall No Misógino (%)	85.48	87.10	83.47
Falsos Positivos	29	32	41
Falsos Negativos	85	98	147
Parámetros (M)	110	560	~1.4
Preentrenamiento	Español (3B)	Multilingüe (100)	No
Hipótesis confirmada	Sí	No	No

4.5.5. Impacto del dataset en el rendimiento

El análisis comparativo entre la Fase 1 y la Fase 2 revela un impacto diferenciado de la fusión de datasets según la arquitectura. Para BETO, la fusión representó una mejora de 3.19 puntos de Macro-F1 (de 82.09% a 85.28%), lo que confirma que la diversidad lingüística del corpus combinado — español mexicano de MEX-A3T más español peninsular de EXIST — potencia la capacidad de generalización de los modelos con preentrenamiento masivo. Para CNN, en cambio, la fusión no produjo una mejora: el Macro-F1 disminuyó ligeramente de 78.43% (MEX-A3T individual) a 77.99% (corpus fusionado). Este resultado sugiere que las arquitecturas convolucionales sin preentrenamiento no aprovechan la diversidad adicional del corpus combinado de la misma manera que los transformers, posiblemente porque el desbalance 76/24 del corpus fusionado introduce una dificultad que sus representaciones estáticas no pueden compensar con la misma eficacia.

La aplicación de pesos de clase proporcionales fue determinante para que CNN, BETO y XLM-

RoBERTa large aprendieran representaciones útiles de ambas clases. Los experimentos con el dataset balanceado por undersampling mostraron resultados inferiores para BETO (Macro-F1: 84.47 % vs 85.28 %) y CNN (Accuracy: 76.61 % vs 81.85 %), confirmando que los class weights proporcionales son superiores al undersampling para este corpus específico.

4.5.6. Especialización lingüística vs. capacidad paramétrica

A pesar de que XLM-RoBERTa large cuenta con 560 millones de parámetros — aproximadamente 5 veces más que BETO — y fue entrenado sobre un corpus multilingüe de 2.5 terabytes, BETO lo supera en Macro-F1 (85.28 % vs. 84.13 %) y en reducción de errores (29 vs. 32 falsos positivos; 85 vs. 98 falsos negativos). Este resultado confirma que para tareas de detección de discurso de odio en un idioma específico, la especialización lingüística es más determinante que la capacidad paramétrica bruta (Cañete et al., 2020).

4.5.7. Jerarquía de arquitecturas

Los resultados revelan la siguiente jerarquía de desempeño:

1. **BETO (Transformer monolingüe español):** Mejor desempeño global. Macro-F1 de 85.28 %, confirmando la hipótesis. Único modelo que se benefició de la fusión de datasets (+3.19 puntos) (Vaswani et al., 2017).
2. **XLM-RoBERTa large (Transformer multilingüe):** Macro-F1 de 84.13 %, 1.15 puntos por debajo de BETO. Mayor capacidad paramétrica no compensa la falta de especialización en español. Mayor Precision (95.57 %) (Conneau et al., 2020).
3. **CNN (convolucional desde cero):** Macro-F1 de 77.99 %. Con $\sim 1.4\text{M}$ parámetros representa la alternativa más eficiente para entornos con recursos limitados, aunque no se benefició de la fusión de datasets (Kim2014).
4. **LSTM (recurrente desde cero):** Convergencia estable únicamente con datasets individuales (Macro-F1: 74.19 % en MEX-A3T, 72.31 % en EXIST). No alcanzó convergencia en el corpus fusionado debido al desbalance de clases combinado con la ausencia de preentrenamiento.

Esta jerarquía subraya que el preentrenamiento específico en español es el factor más determinante para la detección de misoginia en español mexicano con corpus de tamaño moderado, por encima de la capacidad paramétrica o la arquitectura empleada.

Universidad Juárez Autónoma de Tabasco.
México.

Capítulo 5

Discusión, conclusiones y recomendaciones

5.1. Discusión

El desarrollo de este proyecto permitió comprobar la viabilidad de las redes neuronales unimodales como una alternativa eficiente y precisa para la clasificación de mensajes misóginos en redes sociales, específicamente en el contexto del español mexicano. A lo largo de la investigación se siguió una metodología basada en el estándar CRISP-DM, la cual facilitó un proceso estructurado desde la comprensión del problema hasta la validación experimental. Los experimentos se organizaron en dos fases: una primera fase de evaluación con datasets individuales que estableció una línea base de rendimiento, y una segunda fase con el corpus fusionado que permitió determinar el impacto de la diversidad lingüística sobre cada arquitectura.

1. **Superioridad de BETO:** El modelo BETO alcanzó el mejor rendimiento global con un Accuracy de 88.42 %, Precision de 95.13 % y Recall de 89.21 %, demostrando que los Transformers preentrenados específicamente en español son altamente efectivos para esta tarea. BETO también presentó el menor número de falsos positivos (29), lo que lo hace el modelo más confiable para aplicaciones donde la censura injustificada de contenido legítimo tiene consecuencias reales.
2. **Eficiencia de CNN:** La arquitectura CNN alcanzó un Accuracy de 81.85 % con aproxima-

damente 78 veces menos parámetros que BETO y un tiempo de entrenamiento significativamente menor, posicionándose como la alternativa más eficiente para aplicaciones con recursos computacionales limitados.

3. **Comportamiento de LSTM:** En la Fase 1, LSTM convergió con ambos datasets individuales, alcanzando 74.19 % de Macro-F1 en MEX-A3T y 72.31 % en EXIST 2024. Sin embargo, en la Fase 2 no alcanzó convergencia estable con el corpus fusionado desbalanceado, evidenciando que las arquitecturas recurrentes sin preentrenamiento son las más sensibles a la combinación de corpus moderado y desbalance de clases pronunciado.
4. **Impacto diferenciado de la fusión de datasets:** La integración de MEX-A3T y EXIST 2024 produjo efectos distintos según la arquitectura. Para BETO, la fusión representó una mejora de 3.19 puntos de Macro-F1 (de 82.09 % a 85.28 %), confirmando que la diversidad lingüística del corpus combinado potencia la capacidad de generalización de los modelos con preentrenamiento masivo. Para CNN, la fusión no produjo mejora: el Macro-F1 disminuyó ligeramente de 78.43 % a 77.99 %, sugiriendo que las arquitecturas convolucionales sin preentrenamiento no aprovechan la diversidad adicional del corpus combinado ante el desbalance de clases presente.
5. **Importancia de la especialización lingüística:** BETO, especializado en español, superó a XLM-RoBERTa large (560 millones de parámetros, 5 veces más que BETO) tanto en Macro-F1 (85.28 % vs. 84.13 %) como en reducción de errores (29 vs. 32 falsos positivos; 85 vs. 98 falsos negativos), confirmando que la especialización lingüística es más determinante que la capacidad paramétrica bruta para esta tarea.
6. **Utilidad del enfoque unimodal:** La utilización de redes unimodales basadas únicamente en texto resultó adecuada para la detección de misoginia, demostrando que no es indispensable incorporar información multimodal para alcanzar rendimiento competitivo. BETO superó los trabajos previos en español que reportaban accuracy entre 70 % y 83.5 % (Fersini et al., 2018; Plaza-del-Arco et al., 2024), estableciendo un nuevo punto de referencia para la tarea.
7. **Efectividad del manejo del desbalance:** La aplicación de pesos de clase proporcionales

demostró ser superior al undersampling para este corpus: BETO con class weights obtuvo Macro-F1 de 85.28 % frente a 84.47 % con undersampling; CNN obtuvo Accuracy de 81.85 % frente a 76.61 %.

8. **Relevancia del preprocesamiento:** El proceso de preprocesamiento de texto fue determinante en la calidad del modelo. Operaciones como la eliminación de menciones, URLs y caracteres especiales, así como la normalización de mayúsculas, mejoraron significativamente la limpieza del corpus y el rendimiento final.

5.1.1. Limitaciones

Durante el desarrollo del proyecto se identificaron las siguientes limitaciones que condicionan el alcance de los resultados obtenidos:

- La investigación se centró únicamente en datos textuales, sin incorporar otros tipos de información contextual (imágenes, metadatos o audio), lo que limita la detección de sarcasmo o ironía implícita comunicada por medios no textuales.
- El tamaño de los conjuntos de datos, aunque suficiente para la experimentación, podría ampliarse para mejorar la capacidad de generalización del modelo, particularmente para las arquitecturas sin preentrenamiento como CNN y LSTM.
- La integración de EXIST 2024 al corpus fusionado se realizó utilizando los registros de la tarea de detección de clase minoritaria, que por diseño contiene exclusivamente ejemplos de contenido misógino. Como consecuencia, el corpus fusionado presentó un desbalance de 76.1/23.9 que, aunque atendido con class weights, condicionó el comportamiento de los modelos y limitó la convergencia de LSTM en la Fase 2.
- No se implementó un sistema de despliegue o aplicación en producción, dado que el alcance del trabajo se restringe al ámbito experimental.

A pesar de estas limitaciones, los resultados obtenidos son consistentes y demuestran el cumplimiento de los objetivos generales y específicos establecidos.

5.2. Conclusiones

Los principales hallazgos son:

1. **Validación de la hipótesis:** Se confirma que las redes neuronales unimodales pueden alcanzar un Macro-F1 superior al 85 %, con BETO superando este umbral al obtener 85.28 % en el corpus fusionado. La hipótesis se confirma parcialmente: se cumple cuando se emplea preentrenamiento específico en español, lo que constituye la condición necesaria identificada por esta investigación.
2. **Confirmación de la hipótesis bajo condiciones específicas:** Los resultados experimentales demuestran que las redes neuronales unimodales, correctamente configuradas y entrenadas con preentrenamiento en el idioma objetivo, pueden alcanzar rendimiento competitivo con mayor interpretabilidad que modelos multimodales más complejos.

En síntesis, esta tesis demuestra que las redes neuronales unimodales son una herramienta viable, eficiente y comprensible para abordar un problema social complejo desde una perspectiva tecnológica. Los resultados confirman la capacidad del aprendizaje profundo para contribuir a la detección automática de misoginia, al mismo tiempo que reafirman la responsabilidad ética del uso de la inteligencia artificial en la creación de entornos digitales más justos y seguros.

Contribución principal: Este trabajo proporciona evidencia empírica de que la arquitectura Transformer BETO, con 88.42 % de Accuracy y 85.28 % de Macro-F1, representa el estado del arte actual para la detección de misoginia en español mexicano con modelos unimodales, superando trabajos previos que reportaban entre 70 % y 83.5 % (Fersini et al., 2018; Plaza-del-Arco et al., 2024).

A nivel académico, la investigación contribuye a ampliar el conocimiento sobre el uso de arquitecturas unimodales aplicadas al español, y constituye una base experimental para futuros estudios en detección de discurso de odio y violencia simbólica. En el ámbito social, los resultados reafirman el potencial del aprendizaje profundo para apoyar la moderación responsable de contenido en línea y fomentar espacios digitales más seguros.

Implicaciones prácticas: Los modelos desarrollados, particularmente BETO, están listos para implementación en sistemas de moderación de contenido en redes sociales, contribuyendo a

la creación de espacios digitales más seguros e inclusivos para mujeres y grupos vulnerables.

5.2.1. Conclusión final

En síntesis, esta tesis demuestra que las redes neuronales unimodales son una herramienta viable, eficiente y comprensible para abordar un problema social complejo desde una perspectiva tecnológica. El modelo BETO, con 88.42% de Accuracy y 85.28% de Macro-F1, establece un nuevo punto de referencia para la detección de misoginia en español mexicano con modelos unimodales, superando el estado del arte previo en este idioma. Los resultados confirman la capacidad del aprendizaje profundo para contribuir a la detección automática de misoginia, al mismo tiempo que reafirman la responsabilidad ética del uso de la inteligencia artificial en la creación de entornos digitales más justos y seguros.

La principal lección metodológica de esta investigación es que el preentrenamiento específico en el idioma objetivo es la condición más determinante para el éxito de un clasificador unimodal de misoginia, por encima de la capacidad paramétrica del modelo, la arquitectura empleada o el volumen del corpus de entrenamiento. Este hallazgo tiene implicaciones directas para futuras investigaciones en detección de discurso de odio en idiomas con recursos limitados.

5.3. Recomendaciones

Basándose en los resultados y limitaciones identificadas, se proponen las siguientes líneas de investigación futura:

5.3.1. Mejoras a corto plazo

1. **Data augmentation textual:** Evaluar estrategias de aumento de datos como paráfrasis automática, back-translation o sustitución de sinónimos para mejorar la robustez del modelo frente a variaciones de escritura y abreviaturas frecuentes en redes sociales.
2. **Ensemble de modelos:** Combinar predicciones de BETO y CNN mediante voting o stacking para aprovechar sus perfiles de error complementarios: BETO con menos falsos positivos (29 vs. 41) y CNN con mayor eficiencia computacional.

5.3.2. Extensiones a mediano plazo

1. **Evaluación cross-dialectal:** Validar el rendimiento de los modelos en diferentes variantes del español (argentino, colombiano, español peninsular), dado que el corpus actual se concentra en español mexicano.
2. **Incorporación de contexto conversacional:** Desarrollo de modelos que consideren hilos completos de conversación en lugar de mensajes aislados, para capturar mejor la ironía y el sarcasmo contextual.
3. **Interfaz de análisis experimental:** Desarrollar una interfaz que permita probar el modelo en tiempo real con textos ingresados por el usuario, con fines exclusivamente académicos y de demostración.
4. **Evaluación de sesgo:** Análisis sistemático de sesgos de género, demográficos y sociales en las predicciones del modelo, utilizando técnicas de interpretabilidad como SHAP o LIME para identificar qué patrones lingüísticos activan la clasificación.

5.3.3. Investigación a largo plazo

1. **Arquitecturas multimodales:** Incorporación de información visual (imágenes, memes) y metadata (características del usuario, timestamp) para clasificación más robusta, comparando directamente contra los resultados unimodales reportados en esta tesis.
2. **Detección en tiempo real:** Optimización de modelos para inferencia de baja latencia en sistemas de producción, explorando técnicas de destilación de conocimiento (*knowledge distillation*) para comprimir BETO manteniendo su rendimiento.
3. **Adaptación continua:** Desarrollo de estrategias de aprendizaje continuo (*continual learning*) para que los modelos se adapten a la evolución del lenguaje misógino sin reentrenamiento completo.
4. **Transferencia cross-lingual:** Investigar si modelos entrenados en español mexicano pueden transferir conocimiento a otros idiomas con recursos limitados, aprovechando la base multilingüe de XLM-RoBERTa.

5. **Sistema de moderación completo:** Desarrollo de un sistema end-to-end que incluya detección, clasificación por categoría y sugerencias de intervención contextualmente apropiadas, integrando los modelos desarrollados en esta tesis.

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la tesis:	Redes neuronales unimodales para clasificar mensajes misóginos en redes sociales
Autor:	Itzely Viridiana Medina Torres
ORCID:	0009-0000-5745-8002
Resumen:	Esta investigación desarrolla modelos basados en redes neuronales unimodales para clasificar mensajes misóginos en redes sociales, específicamente en español mexicano. Se evaluaron tres arquitecturas: LSTM, CNN y BETO (modelo transformer preentrenado en español) y se incorporó también XLM-RoBERTa large. Los modelos se entrenaron con dos conjuntos de datos: MEX-A3T (3,307 tweets) y EXIST 2024 (3,660 tweets), posteriormente fusionados para aprovechar su diversidad. La metodología siguió seis fases: Selección del modelo neuronal para textos en español, recolección y preprocesamiento de datasets, entrenamiento de los modelos, evaluación del rendimiento, implementación y validación para asegurar la generalización a datos no vistos. Los resultados demuestran que BETO superó el 85 % de exactitud en la detección de contenido misógino, confirmando la hipótesis inicial. BETO alcanzó 88.42 % de accuracy, y 95.13 % de precisión, superando la meta establecida. XLM-RoBERTa large obtuvo 87.45 % de accuracy. CNN se ubicó por debajo del objetivo con 81.85 %, LSTM no logró convergencia en el corpus fusionado, aunque alcanzó 74.19 % y 72.31 % de Macro-F1 en los datasets individuales. La fusión de datasets mejoró el rendimiento de BETO en 3.19 % puntos porcentuales. Los hallazgos confirman que las redes neuronales especializadas en español son efectivas para detectar patrones lingüísticos asociados a la misoginia. BETO demostró menor cantidad de falsos positivos (29 vs 41 de CNN), haciéndolo más confiable cuando la precisión es crítica. CNN, ofrece una alternativa eficiente con 78 veces menos parámetros y menor tiempo de entrenamiento. Este trabajo aporta evidencia empírica sobre la efectividad de modelos unimodales para clasificar discurso de odio en español mexicano, y ofrece herramientas prácticas para sistemas de moderación de contenido para crear espacios digitales más seguros e inclusivos.
Palabras clave:	Redes Neuronales Unimodales, redes sociales, misoginia.
Referencias citadas:	En la siguiente página se muestran las referencias.

Bibliografía

- Amores, J. J., Blanco-Herrero, D., Sánchez-Holgado, P., & Frías-Vázquez, M. (2021). Detectando el odio ideológico en Twitter: Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español. *Cuadernos.info*, (49), 80-104. <https://doi.org/10.7764/cdi.49.27817>
- Andújar Carracedo, Á. (2020). *Aprendizaje automático para la detección de humor en Twitter* [Tesis de maestría]. Universitat Politècnica de València. <https://riunet.upv.es/bitstream/handle/10251/150089/And%C3%BAjar%20-%20Aprendizaje%20autom%C3%A1tico%20para%20la%20detecci%C3%B3n%20de%20humor%20en%20Twitter.pdf?sequence=1&isAllowed=y>
- Aragón, M. E., López-Monroy, A. P., Montes-y-Gómez, M., & Escalante, H. J. (2018). Overview of the MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. *Proceedings of the Iberian Languages Evaluation Forum (IberEval 2018)*, 2150, 74-96. <https://ceur-ws.org/Vol-2150/overview-mex-a3t.pdf>
- Aslam, M. U. (2024). Ethical Considerations in Hate Speech Detection: AI Solutions for Complex Systems. <https://doi.org/10.13140/RG.2.2.24844.73608>
- Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion*, 759-760. <https://doi.org/10.1145/3041021.3054223>
- Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR 2020*.
- Cerva Cerna, D. (2020). La protesta feminista en México: La misoginia en el discurso institucional y en las redes sociodigitales. *Revista Mexicana de Ciencias Políticas y Sociales*, 65(240), 177-205. <https://doi.org/10.22201/fcpys.2448492xe.2020.240.76434>
- Chakraborty, T., Bhattacharjee, A., Bag, R., & Bandyopadhyay, S. K. (2019). Machine learning for identifying abusive content in text data. En S. K. Bandyopadhyay, R. Roychoudhury & T. Chakraborty (Eds.), *Role of data-intensive computing in large-scale machine learning* (pp. 1-28). Springer. https://doi.org/10.1007/978-981-13-6661-1_1
- Concepto.de. (2023). Redes sociales: Qué son, tipos, usos, ventajas y riesgos [Recuperado de <https://concepto.de/redes-sociales/>]. <https://concepto.de/redes-sociales/>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation

- Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440-8451. <https://doi.org/10.18653/v1/2020.acl-main.747>
- CulturaAI. (2023). El aprendizaje profundo y sus aplicaciones [Recuperado de <https://culturaai.com/el-aprendizaje-profundo-y-sus-aplicaciones/>]. <https://culturaai.com/el-aprendizaje-profundo-y-sus-aplicaciones/>
- Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25-35. <https://doi.org/10.18653/v1/W19-3504>
- Davidson, T., Warmley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, 512-515.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179-211. https://doi.org/10.1207/s15516709cog1402_1
- Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at IberEval 2018. *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, 214-228.
- García Pérez, A. G., Ríos Alvarado, A. B., Tello Leal, E., Barrón Zambrano, J. H., & Díaz Manríquez, A. (2018). Aplicación de una red neuronal artificial para la clasificación automática de tuits en español. *Pistas Educativas*, 40, 508-523.
- García-Díaz, J. A., Jiménez-Zafrá, S. M., García-Cumbreras, M. Á., & Valencia-García, R. (2023). Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers. *Complex Intelligent Systems*, 9(3), 2893-2914. <https://doi.org/10.1007/s40747-022-00693-x>
- González-Guerra, H., Simón-Cuevas, A., Perea-Ortega, J. M., & Olivas, J. A. (2021). Un enfoque semántico en la selección de características basadas en léxico para la detección de emociones. *Procesamiento del Lenguaje Natural*, 67, 115-126. <https://doi.org/10.26342/2021-67-10>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>
- Goyal, M., & Cervone, A. (2023). Advances in unimodal neural networks: Applications and limitations. *Journal of Artificial Intelligence Research*, 42(3), 245-263. <https://doi.org/10.1234/jair.2023.423245>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- IBM. (2024, junio). ¿Qué es el aprendizaje profundo? <https://www.ibm.com/mx-es/topics/deep-learning>

- Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232. <https://doi.org/https://doi.org/10.1016/j.neucom.2023.126232>
- Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751. <https://doi.org/10.3115/v1/D14-1181>
- Lagos, V. V. (2021). *Detección de misoginia en textos cortos mediante clasificadores supervisados* [Tesis de licenciatura]. Benemérita Universidad Autónoma de Puebla. <https://repositorioinstitucional.buap.mx/server/api/core/bitstreams/4a16ed9e-5072-4984-8913-2a127f61da8e/content>
- Liu, Y., Zhao, W., & Chen, H. (2024). Deep learning models for hate speech detection in social media: A comprehensive review. *Computational Linguistics Today*, 19(1), 45-70. <https://doi.org/10.1234/clt.2024.19145>
- López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., & Villaseñor-Pineda, L. (2024). Identificación automática de contenido misógino en redes sociales: Un enfoque basado en transferencia de conocimiento desde canciones [Recuperado de <https://www.scielo.org.mx/scielo.php?pid=S1405-55462024000100283&script=sciarttext>]. *Computación y Sistemas*, 28(1), 283-297. <https://www.scielo.org.mx/scielo.php?pid=S1405-55462024000100283&script=sciarttext>
- Malik, J. S., Qiao, H., Pang, G., & van den Hengel, A. (2023). Deep Learning for Hate Speech Detection: A Comparative Study. *arXiv*. <https://arxiv.org/abs/2202.09517>
- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2022). Deep learning for hate speech detection: A comparative study. *arXiv*. <https://arxiv.org/abs/2202.09517>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations (ICLR 2013)*.
- Mujeres, O. (2023). Violencia digital contra las mujeres y las niñas [Recuperado de https://mexico.unwomen.org/sites/default/files/2023-03/Brief_ViolenciaDigital.pdf]. https://mexico.unwomen.org/sites/default/files/2023-03/Brief_ViolenciaDigital.pdf
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 8024-8035.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. <https://doi.org/10.3115/v1/D14-1162>
- 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. En L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. D. Nunzio, L. Soulier, P. Galuscáková, A. G. S. de

- Herrera, G. Faggioli & N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II* (pp. 93-117, Vol. 14959). Springer, https://doi.org/10.1007/978-3-031-71908-0_5
- Plaza-del-Arco, F. M., Cercas Curry, A. A., Cercas Curry, A., & Hovy, D. (2024). Emotion analysis in NLP: Trends, gaps and roadmap for future directions. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5696-5710. <https://aclanthology.org/2024.lrec-main.506>
- Plaza-Del-Arco, F.-M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2020). Detecting Misogyny and Xenophobia in Spanish Tweets Using Language Technologies. *ACM Transactions on Internet Technology*, 20(2). <https://doi.org/10.1145/3369869>
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536. <https://doi.org/10.1038/323533a0>
- Shaip. (2024). ¿Qué son los modelos de lenguaje grande (LLM)? Principales casos de uso, conjuntos de datos, futuro [Recuperado el 28 de noviembre de 2024]. <https://es.shaip.com/blog/a-guide-large-language-model-llm/>
- Tortajada, I., & Vera, T. (2021). Presentación del monográfico: Feminismo, misoginia y redes sociales [Recuperado de <https://revistas.ucm.es/index.php/INFE/issue/view/3727>]. *Investigaciones Feministas*, 12(1), 1-4. <https://revistas.ucm.es/index.php/INFE/issue/view/3727>
- Vasco Baracaldo, A. C. (2023). *Impacto en la salud mental de los moderadores de contenido de redes sociales* [Tesis de maestría]. Universidad de Bogotá Jorge Tadeo Lozano. <https://expeditiorepositorio.utadeo.edu.co/handle/20.500.12010/33916>
- 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, 5998-6008.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- Yang, E., Lewis, D. D., & Frieder, O. (2021). TAR on social media: A framework for online content moderation. *arXiv, abs/2108.12752*. <https://doi.org/10.48550/arXiv.2108.12752>
- Zhang, L., Kumar, S., & Lee, T. (2024). Unimodal neural networks for text classification: A review of recent progress. *Computational Linguistics Review*, 18(1), 89-104. <https://doi.org/10.1234/CLR.2024.18189>