MÁSTER EN FORMACIÓN PERMANENTE EN INTELIGENCIA ARTIFICIAL

TRANSFORMERS PARA LA DETECCIÓN DE SEXISMO EN TWEETS CON ETIQUETADO AMBIGUO

Presentado por:

LUISA FERNANDA RIASCOS LULIGO

Dirigido por:

VICTOR MANUEL YESTE MORENO

CURSO ACADÉMICO 2024/2025

Resumen

Este Trabajo de Fin de Máster aborda el desafío de la detección automática de sexismo en redes sociales, un problema caracterizado por la ambigüedad del lenguaje y la subjetividad de la interpretación. En el marco de la competición internacional EXIST 2025, se diseña, implementa y evalúa una novedosa arquitectura híbrida que fusiona dos paradigmas de la inteligencia artificial. Por un lado, se optimiza un modelo Transformer especializado (XLM-RoBERTa-Large) mediante ajuste fino avanzado y aprendizaje multi-tarea para la clasificación precisa de patrones explícitos. Por otro lado, se desarrolla un sistema de inferencia dinámica basado en un Modelo de Lenguaje Grande (Mistral) potenciado por una estrategia de Generación Aumentada por Recuperación (RAG) multi-etapa, diseñada para resolver casos ambiguos mediante la inyección de contexto en tiempo real. Los resultados experimentales, evaluados sobre el conjunto de datos de desarrollo, demuestran un claro trade-off: el Transformer especializado sobresale en la clasificación binaria (F1-Score de 0.81), mientras que el sistema híbrido es significativamente más robusto en la compleja tarea de clasificación de intención (ICM-Norm de 0.46 vs. 0.00 del Transformer). La contribución principal es una arquitectura sinérgica y explicable que combina la precisión del especialista con la capacidad de razonamiento del generalista, representando una solución más completa y resiliente para la moderación de contenido.

Palabras clave: Detección de Sexismo, Procesamiento del Lenguaje Natural, Modelos Transformer, Generación Aumentada por Recuperación (RAG), Aprendizaje con Desacuerdo (LeWiDi), Inteligencia Artificial Híbrida, Explicabilidad (XAI).

Abstract

This Master's Thesis addresses the challenge of automatic sexism detection on social media, a problem characterized by language ambiguity and subjective interpretation. Within the framework of the EXIST 2025 international challenge, a novel hybrid architecture that merges two artificial intelligence paradigms is designed, implemented, and evaluated. On one hand, a specialized Transformer model (XLM-RoBERTa-Large) is optimized through advanced fine-tuning and multi-task learning for the precise classification of explicit patterns. On the other hand, a dynamic inference system based on a Large Language Model (Mistral) is developed, powered by a multi-stage Retrieval-Augmented Generation (RAG) strategy designed to resolve ambiguous cases by injecting real-time context. Experimental results, evaluated on the development dataset, reveal a clear trade-off: the specialized Transformer excels at binary classification (F1-Score of 0.81), while the hybrid system is significantly more robust in the complex intention classification task (ICM-Norm of 0.46 vs. the Transformer's 0.00). The main contribution is a synergistic and explainable architecture that combines the specialist's precision with the generalist's reasoning capabilities, representing a more complete and resilient solution for content moderation.

Keywords: Sexism Detection, Natural Language Processing, Transformer Models, Retrieval-Augmented Generation (RAG), Learning with Disagreement (LeWiDi), Hybrid Artificial Intelligence, Explainable AI (XAI).

Contenido

1. Introducción

1.1 Contexto y problemática social del sexismo en redes

El sexismo ha estado presente a lo largo de la historia como un fenómeno estructural que perpetúa desigualdades entre géneros, afectando especialmente a mujeres. Esta desigualdad se manifiesta en prácticas culturales, sociales y simbólicas, muchas veces naturalizadas. Uno de los vehículos más poderosos para su transmisión ha sido, y continúa siendo, el lenguaje. A través de expresiones, refranes, chistes o estereotipos lingüísticos, se refuerzan roles tradicionales, se invisibiliza la discriminación y se legitima la subordinación de ciertos grupos. En muchos casos, este lenguaje sexista se ha interiorizado a tal punto que es reproducido sin conciencia de su carga simbólica.

Con la irrupción y expansión global de las redes sociales, este fenómeno no solo ha encontrado un nuevo canal de difusión, sino que ha mutado, adaptándose a las particularidades del ecosistema digital. Plataformas como X se han convertido en espacios donde el lenguaje sexista se propaga rápidamente, con un alcance masivo y una visibilidad que trasciende fronteras. A menudo, las personas no identifican ciertas expresiones como sexistas debido a su normalización cultural, lo que dificulta aún más su detección y cuestionamiento.

Esta transición al entorno digital ha introducido nuevas dimensiones al problema, como el anonimato que fomenta ataques más directos, la viralidad que puede generar campañas de acoso coordinado y la normalización a través de nuevos formatos como los memes, haciendo su detección aún más compleja por lo cual se hace urgente el desarrollo de herramientas automatizadas que permitan identificar estos discursos a gran escala, de forma eficiente y sensible a sus matices.

1.2 El Desafío Técnico: Detección Automática y el Reto EXIST 2025

La erradicación del sexismo en las plataformas digitales representa un desafío técnico de gran envergadura. El desarrollo de sistemas automáticos de detección es una necesidad imperante, dado que la moderación manual de contenidos es inviable debido al volumen y la velocidad con la que se genera la información. La complejidad de esta tarea se deriva directamente de la naturaleza multifacética del sexismo, un espectro que abarca desde la violencia explícita hasta formas implícitas y normalizadas de discriminación. El principal obstáculo es la dependencia contextual del lenguaje. "El principal obstáculo es la dependencia contextual del lenguaje. Expresiones sutiles como el sarcasmo o el humor pueden ocultar la intención sexista. En este sentido (Ruiz et al., 2024) advierten que «los tuits a menudo usan jerga, sarcasmo o lenguaje codificado que puede ocultar la intención sexista»

A menudo, para entender bien un texto ambiguo, no basta con lo que está escrito. Es necesario tener en cuenta el contexto exterior, como lo que está pasando en el mundo, la cultura o las modas. Además, La importancia de analizar el sexismo en su contexto lingüístico es evidente, ya que sus formas varían notablemente. Por ejemplo, (Rodríguez-Sánchez et al., 2024) encontraron que los textos en español suelen reflejar el neosexismo, a diferencia de los textos en inglés, donde las connotaciones son mayoritariamente sexuales. Para analizar un fenómeno tan complejo como el sexismo online, no basta con identificarlo de forma genérica. Es crucial desglosarlo en sus componentes fundamentales para poder medirlo y combatirlo de manera efectiva. En este sentido, la creación de una taxonomía robusta es un paso metodológico indispensable que permite pasar de la simple detección a una comprensión más profunda del problema. Es aquí donde campañas como EXIST han sido determinantes, pues, como señala (Plaza et al., n.d.) han refinado la taxonomía del problema, incluyendo categorías como ideología y desigualdad, estereotipos y dominación, o violencia sexual.

La relevancia de esta clasificación es doble. Por un lado, visibiliza que el sexismo no es un acto único, sino un espectro que abarca desde la perpetuación de creencias dañinas (ideología y desigualdad) y la imposición de roles (estereotipos y dominación) hasta sus formas más explícitas y violentas. Por otro lado, ofrece a este trabajo un marco de análisis concreto que permite no solo detectar la presencia de sexismo, sino también clasificar su naturaleza, lo cual es fundamental para evaluar su impacto y las estrategias para contrarrestarlo.

Uno de los mayores desafíos al crear conjuntos de datos para tareas subjetivas como la detección de sexismo es la inevitable discrepancia entre los anotadores humanos (Plaza et al., n.d.). Lo que una persona considera claramente ofensivo, otra puede interpretarlo como ambiguo. Los enfoques de etiquetado tradicionales suelen tratar este desacuerdo como un "ruido" que debe ser eliminado para forzar un consenso en una única etiqueta de verdad ("ground truth").

Frente a esta limitación, que puede simplificar en exceso el fenómeno y descartar información valiosa sobre los matices del lenguaje, han surgido nuevos paradigmas metodológicos. Un ejemplo destacado es el enfoque adoptado por la competición internacional EXIST, la cual basa su proceso de evaluación en el paradigma.

Learning With Disagreement (LeWiDi). Según se detalla en las guías oficiales del taller (Plaza et al., n.d.) , en lugar de reducir las múltiples anotaciones a una única etiqueta "dura" (por ejemplo, mediante un voto mayoritario), el enfoque "soft" utiliza el conjunto completo de anotaciones humanas, con toda su variabilidad, como la verdad de referencia. Esto se consigue mediante el uso de etiquetas probabilísticas que reflejan la proporción de acuerdo entre los distintos anotadores para cada categoría.

Al hacer esto, el desacuerdo deja de ser un obstáculo para convertirse en una característica informativa que enseña a los modelos sobre la ambigüedad inherente al

sexismo. Esta metodología, por tanto, no solo resulta más fiel a la realidad de la percepción humana, sino que fomenta el desarrollo de sistemas de inteligencia artificial más robustos y capaces de navegar la complejidad del lenguaje subjetivo.

La respuesta de la comunidad investigadora se ha centrado en el ajuste fino (fine-tuning) de modelos Transformer. Si bien son eficaces, su naturaleza de "caja negra" presenta un reto para la transparencia.

Para superar estas limitaciones, la investigación actual explora nuevas fronteras como el aprendizaje multitarea y arquitecturas que integran conocimiento externo, como la Generación Aumentada por Recuperación (RAG). Paralelamente, el auge de los Modelos de Lenguaje Grandes (LLMs) ha introducido nuevos paradigmas, aunque su aplicación no está exenta de desafíos. Un estudio de (Khreich & Doughman, 2025)) encontró que "su rendimiento en tareas de clasificación de nicho puede ser inferior al de modelos más pequeños y especializados que han sido ajustados finamente en datos específicos de la tarea".

El presente Trabajo de Fin de Máster se inscribe en esta frontera, proponiendo y evaluando dos arquitecturas computacionales: un modelo Transformer y un novedoso sistema híbrido basado en Transformers y LLM con RAG. El objetivo es abordar los retos de la competición y realizar un análisis comparativo profundo en términos de precisión, interpretabilidad y eficiencia, para contribuir al desarrollo de herramientas más robustas y conscientes del contexto para la moderación de contenido en línea.

1.3 Justificación y Propuesta de Valor

La proliferación de contenido sexista en las redes sociales no solo perpetúa estereotipos dañinos, sino que también crea entornos digitales hostiles que pueden silenciar voces y afectar negativamente el bienestar de los usuarios. La detección automática de este tipo de contenido, por tanto, no es solo un desafío técnico, sino una necesidad social para fomentar espacios en línea más seguros e inclusivos. Si bien los modelos Transformer han

demostrado ser la columna vertebral de la clasificación de texto moderna, su eficacia está intrínsecamente ligada a su capacidad para comprender los matices del lenguaje. Este proyecto se justifica en la necesidad de ir más allá de los enfoques de clasificación convencionales para abordar las limitaciones clave de los sistemas actuales: la dependencia del contexto y la interpretabilidad.

La propuesta consiste en entrenar y evaluar dos paradigmas avanzados de la inteligencia artificial. Por un lado, se desarrollará un modelo Transformer adaptado a cada tarea, aplicando técnicas avanzadas de fine-tuning, congelamiento selectivo de capas y ajuste de umbrales. Este enfoque se basa en entrenar un modelo para realizar tareas relacionadas simultáneamente (como identificar sexismo y categorizar la intención) le obliga a aprender representaciones más ricas y generalizables. Como se ha observado en investigaciones previas, "el aprendizaje multitarea puede llevar a una mejor regularización y a una mejor capacidad de generalización del modelo" Haga clic o pulse aquí para escribir texto. (Y. Zhang & Yang, 2021) lo que es crucial para una tarea tan subjetiva como la detección de sexismo. La evaluación de este modelo se realizará de manera robusta mediante herramientas de análisis de rendimiento como PyEvALL.

Por otro lado, se diseñará y evaluará un sistema de Generación Aumentada por Recuperación (RAG). Este enfoque es particularmente innovador para tareas de clasificación, ya que dota al modelo de la capacidad de consultar una base de conocimiento externa antes de tomar una decisión. Esta arquitectura es una respuesta directa al problema del conocimiento estático en los LLMs, que a menudo carecen de contexto sobre eventos recientes o memes culturales específicos. La justificación de este segundo modelo se alinea con la observación de que "la clasificación correcta a menudo depende de información no presente en el texto mismo (por ejemplo, eventos actuales, referencias culturales, tendencias

emergentes)" (Nowakowski et al8, 2025). Al permitir que el modelo "busque" información relevante, se espera reducir los errores causados por la falta de contexto.

La propuesta de valor de este trabajo es doble. Técnicamente, ofrece una comparativa rigurosa entre un enfoque de optimización de la representación interna (multitarea) y un enfoque de enriquecimiento con conocimiento externo (RAG), proporcionando valiosos insights sobre qué estrategia es más efectiva para la detección de sexismo. Socialmente, el objetivo final es construir un sistema preciso y sensible al contexto, que no solo mejore las métricas de clasificación, sino que también contribuya a visibilizar el sexismo en el entorno digital. Al desarrollar modelos más fiables y potencialmente más explicables, este proyecto busca fomentar una mayor conciencia social mediante el uso responsable de la inteligencia artificial, proveyendo herramientas que puedan servir de base para una moderación de contenido más justa y efectiva.

1.4 Enfoque metodológico y Contribuciones Principales

El enfoque metodológico central de este trabajo se basa en la premisa de que la combinación sinérgica de arquitecturas de IA puede superar las limitaciones de los modelos monolíticos. Se busca demostrar que una arquitectura híbrida, que integra un modelo Transformer especializado con un LLM contextualizado mediante RAG, puede alcanzar un rendimiento superior en la detección de sexismo. El objetivo es sacar lo mejor de ambos mundos: la robustez del ajuste fino supervisado para reconocer patrones explícitos y la capacidad del RAG para inyectar contexto en tiempo real para resolver la ambigüedad.

Para alcanzar este objetivo, este Trabajo de Fin de Máster presenta las siguientes contribuciones principales:

- El diseño e implementación de una arquitectura híbrida que combina un clasificador basado en XLM-RoBERTa con un pipeline RAG+LLM. Esta propuesta aborda la dicotomía entre modelos especializados y generalistas.

Mientras que el clasificador ajustado finamente sobre el corpus de EXIST, proporciona una base de alta precisión para patrones recurrentes, el componente RAG-LLM ofrece una capacidad de razonamiento dinámico. Este diseño se alinea con hallazgos recientes que sugieren que los modelos más pequeños y especializados pueden superar a los grandes LLMs en tareas de nicho (Ghorbanpour et al., 2025)

- La construcción de un sistema RAG avanzado que incluye un mecanismo de re-ranking en dos fases. A diferencia de una recuperación vectorial estándar, que puede introducir ruido, nuestro sistema primero recupera un conjunto de candidatos mediante búsqueda de similitud y, en una segunda etapa, emplea un Cross-Encoder para reevaluar y ordenar estos candidatos según su relevancia contextual precisa con respecto a la consulta. Esta técnica de refinamiento es crucial para garantizar que solo la evidencia más pertinente sea proporcionada al LLM, maximizando la calidad de su razonamiento, así como lo sustenta el artículo original (Lewis et al., n.d.) que introdujo RAG, explica que el modelo combina el conocimiento paramétrico (lo aprendido en sus pesos durante el entrenamiento) con el conocimiento no paramétrico (los documentos recuperados). Este conocimiento no paramétrico es explícito y puede ser inspeccionado, lo que ofrece una ventana a la "memoria" que el modelo está utilizando en un momento dado.
- La creación de tres bases de conocimiento vectoriales especializadas a partir de un corpus con anotación múltiple. En lugar de utilizar una base de conocimiento genérica, se ha diseñado un pipeline que procesa el data set de EXIST, interpretando y consolidando el desacuerdo entre anotadores (Learning with Disagreement) para construir un almacén vectorial que no solo contiene ejemplos, sino también las perspectivas asociadas. Esto permite que el sistema RAG

recupere "puntos de vista" diversos, enriqueciendo la capacidad del LLM para navegar la subjetividad de la tarea.

- Una evaluación experimental rigurosa y un análisis ablativo que cuantifica el valor añadido de cada componente. Se llevará a cabo una comparación sistemática del rendimiento del modelo Transformers aislado, el sistema RAG de forma independiente y la arquitectura híbrida completa. Este análisis permitirá establecer un benchmark claro y demostrar empíricamente el impacto de cada innovación propuesta.
- El pipeline de RAG, al exponer los ejemplos recuperados que fundamentan una decisión, ofrece un mecanismo de interpretabilidad inherente. Esta transparencia aborda directamente el problema de la "caja negra" de muchos modelos de IA.

2. Objetivos

2.1 Objetivo general

Diseñar, implementar y validar una arquitectura de NLP híbrida y sinérgica para la detección y caracterización avanzada del sexismo en tuits, en el marco de la competición EXIST 2025. El sistema buscará fusionar la precisión de un modelo Transformers especializado mediante ajuste fino, y la capacidad de inferencia dinámica de un LLM, potenciado por una estrategia de prompting que combina la recuperación de ejemplos en tiempo real (RAG) con la instrucción de razonamiento interno para maximizar el rendimiento en las tres subtareas de la competición.

2.1 Objetivos específicos

- Desarrollar y optimizar un pipeline de clasificación secuencial basado en Transformers como base de clasificación supervisada. Este componente constará de un modelo binario para la identificación y un modelo multiclase para la intención, integrando técnicas de soft labeling para modelar la subjetividad inherente al desacuerdo entre anotadores.
- Diseñar e implementar un sistema de clasificación dinámica basado en un LLM, aplicable a las tres subtareas, que combine dos estrategias de contextualización: una recuperación de información (RAG) para inyectar ejemplos pertinentes del corpus vectorizado en tiempo real y una ingeniería de prompts avanzada que instruya al modelo a complementar la evidencia recuperada con un razonamiento interno.
- Construir y validar la arquitectura híbrida propuesta, fusionando las predicciones del pipeline Transformer con la salida contextualizada del sistema RAG-LLM para generar una clasificación final sinérgica.

- Realizar un análisis comparativo y ablativo riguroso para evaluar el rendimiento de cada componente (pipeline Transformer, sistema RAG) y de la arquitectura híbrida final. Se utilizarán las métricas oficiales de la competición (ICM, F1-score) para cuantificar la mejora de rendimiento.

3. Antecedentes

El estudio de la detección automática de sexismo es un subcampo del Procesamiento del Lenguaje Natural (PLN) que ha ganado una atracción considerable en los últimos años, impulsado tanto por la necesidad social de moderar el contenido en línea como por los rápidos avances en las arquitecturas de inteligencia artificial. Esta sección revisa la literatura académica relevante, contextualizando el problema y las metodologías existentes para establecer una base sólida que justifique las contribuciones de este trabajo.

3.1 El Sexismo Online como Problema de PLN y el Marco EXIST

El sexismo en las redes sociales es un fenómeno complejo que va más allá del lenguaje explícitamente violento. Se manifiesta a través de "expresiones sutiles que involucran comportamientos sexistas implícitos" (Plaza et al., n.d.), como el humor, los estereotipos o la desinformación ideológica. La naturaleza subjetiva de lo que constituye un mensaje sexista presenta un desafío fundamental para los sistemas de IA.

Para abordar esta subjetividad de manera científica, el taller internacional EXIST (sEXism Identification in Social neTworks) se ha establecido como el principal campo de pruebas para la investigación en esta área. Su contribución más significativa ha sido la adopción del paradigma Learning With Disagreement (LeWiDi). En lugar de forzar un consenso en una única "etiqueta de oro", el dataset de EXIST proporciona las anotaciones de múltiples perfiles de anotadores, definidos por ejes demográficos como género, edad, etnia, nivel de estudios y país (Paola Jimenez-Martinez et al., 2024). Este enfoque permite que los modelos "aprendan de la diversidad de perspectivas, haciéndolos mejores para comprender un rango de puntos de vista humanos" (Plaza et al., n.d.), un paso crucial hacia el desarrollo de una IA más equitativa.

3.2 Evolución de las Arquitecturas de Detección

La metodología para la detección de sexismo ha evolucionado drásticamente, pasando de enfoques estadísticos a complejos modelos neuronales profundos.

3.2.1 La Dominancia de los Modelos Transformer

Desde su introducción, la arquitectura Transformer se ha convertido en el estándar de facto. Prácticamente todos los sistemas de alto rendimiento presentados en EXIST se basan en el ajuste fino (*fine-tuning*) de modelos pre-entrenados como BERT, RoBERTa y sus variantes multilingües (ej. XLM-RoBERTa, mDeBERTa). La elección del modelo a menudo se basa en mejoras arquitectónicas específicas; por ejemplo, **mDeBERTa-v3** es una opción recurrente debido a que "mejora a BERT y ROBERTa implementando atención descentralizada y un decodificador de máscara mejorado" (Ruiz et al., 2024), lo que le confiere una mayor capacidad para capturar relaciones complejas en el texto.

3.2.2 El Auge de los Modelos de Lenguaje Grandes (LLMs)

Más recientemente, el surgimiento de los LLMs ha introducido el paradigma del aprendizaje con pocas o ninguna muestra (*few-shot* o *zero-shot learning*). El éxito de esta técnica depende en gran medida de una cuidadosa **ingeniería de prompts**. Investigaciones dentro de EXIST han demostrado que el rendimiento de un LLM mejora a medida que los prompts se refinan, pasando de una simple pregunta a instrucciones que incluyen definiciones claras de las clases (Tavarez-Rodríguez et al., 2024).

Aprendizaje en Contexto (In-Context Learning)

Más allá de las definiciones (zero-shot), la siguiente etapa evolutiva es el aprendizaje con pocas muestras (few-shot learning). En este paradigma, el prompt no solo define la tarea,

sino que también incluye algunos ejemplos ilustrativos. Por ejemplo, para la detección de sexismo, se le proporcionarían al modelo 2 o 3 ejemplos de tuits etiquetados como sexistas y no sexistas. Como demostraron (Brown et al., 2020) en su artículo sobre GPT-3, proporcionar ejemplos dentro del propio contexto del prompt permite al modelo inferir el patrón y la intención de la tarea sin necesidad de reentrenamiento (actualización de sus pesos). El LLM aprende "en contexto" a realizar la tarea específica.

Técnicas de Razonamiento Guiado

El desarrollo más reciente y sofisticado en la ingeniería de prompts se centra en guiar el "proceso de pensamiento" del modelo. La técnica más destacada en esta área es la Cadena de Pensamiento (Chain of Thought - CoT), propuesta por (Wei et al., 2023). En lugar de pedir directamente una respuesta, el prompt se diseña para que el modelo primero "piense paso a paso" y articule su razonamiento antes de llegar a una conclusión.

Aplicado a la detección de sexismo, un prompt con CoT no solo preguntaría si un texto es sexista, sino que instruiría al modelo a:

1. Identificar los actores y el contexto del mensaje.
2. Analizar si se basa en estereotipos de género.
3. Evaluar si el lenguaje denigra o ataca a alguien por su género.
4. Concluir, basándose en los pasos anteriores, si el texto debe ser clasificado como sexista.

Este enfoque no solo mejora drásticamente la precisión en tareas que requieren lógica y razonamiento, sino que también aporta una invaluable interpretabilidad a las decisiones del modelo. Al explicitar su razonamiento, podemos entender por qué el LLM llegó a una conclusión determinada, facilitando la depuración y el refinamiento de los prompts.

Implicaciones y Desafíos Futuros

El énfasis en la ingeniería de prompts subraya una realidad clave de la era de los LLMs: la habilidad para comunicarse de manera efectiva y precisa con la IA se ha convertido en una disciplina en sí misma.

- **Sensibilidad al fraseo**: El rendimiento de un LLM puede variar enormemente ante cambios aparentemente menores en la redacción del prompt. Esta "fragilidad" de los prompts es un área de investigación activa, buscando métodos para crear modelos más robustos y menos sensibles a la sintaxis exacta de la instrucción (Kroeger et al., 2024)
- **Complejidad y coste**: Diseñar prompts sofisticados, especialmente aquellos que emplean CoT o incluyen múltiples ejemplos (few-shot), requiere un esfuerzo considerable y un profundo conocimiento tanto del dominio del problema como del comportamiento del modelo. Además, prompts más largos y complejos aumentan el coste computacional y la latencia de la inferencia.
- **Hacia la automatización de prompts**: Para superar estos desafíos, la investigación se está moviendo hacia la optimización automática de prompts. Técnicas como el Automatic Prompt Engineer (APE) (Zhou et al., 2023) utilizan LLMs para generar y seleccionar los prompts más efectivos para una tarea determinada, buscando automatizar el laborioso proceso de la ingeniería manual.

3.3 Planteamiento Estratégico de la Arquitectura Híbrida

Para mejorar la robustez, las técnicas de ensamblaje son una práctica habitual. Un enfoque exitoso fue el de "combinar las respuestas de siete sistemas diferentes" (Tavarez-Rodríguez et al., 2024), incluyendo tanto Transformers ajustados como LLMs, demostrando que la diversidad de modelos puede llevar a un consenso más preciso.

3.4 Abordando la Ambigüedad: Hacia Sistemas Conscientes del Contexto

El camino hacia sistemas capaces de comprender el contexto es la respuesta a la ambigüedad y al problema intrínseco de la desactualización del conocimiento que afecta a los modelos estáticos. El lenguaje en redes sociales es dinámico y a menudo hace referencia a eventos de actualidad o memes. Un modelo no puede clasificar correctamente un tuit si no comprende el contexto externo al que alude. Esta limitación ha impulsado la investigación hacia arquitecturas que pueden integrar información en tiempo real, como el uso de "agentes de IA que pueden interactuar dinámicamente con su entorno con herramientas" (Nowakowski et al., 2025) motores de búsqueda.

La Generación Aumentada por Recuperación (RAG) emerge como una formalización de esta idea. En el contexto de la clasificación, RAG permite a un LLM consultar una base de conocimiento para recuperar ejemplos o información relevante antes de emitir un juicio.

3.5 La Necesidad de Transparencia: Explicabilidad (XAI)

A medida que los modelos se vuelven más complejos, también se vuelven más opacos. La naturaleza de "caja negra" de los Transformers y LLMs es una preocupación creciente, especialmente en aplicaciones sensibles como la moderación de contenido. La Explicabilidad (XAI) se ha convertido en un campo de investigación crucial, debido a la creciente dependencia de nuestra sociedad en las máquinas inteligentes y la consiguiente demanda de modelos más transparentes e interpretables(Dwivedi et al., 2023). Un sistema que no solo clasifica, sino que también puede justificar su decisión, es inherentemente más fiable y auditable. El enfoque RAG, al exponer los ejemplos recuperados que informan la decisión del LLM, ofrece de forma natural un grado de interpretabilidad que los modelos de ajuste fino tradicionales no poseen.

4. Marco teórico

El presente trabajo establece los fundamentos conceptuales y técnicos que sustentan la investigación sobre la detección automática de sexismo. Su objetivo es contextualizar el estudio dentro de la literatura científica existente, definir los conceptos clave y justificar las decisiones metodológicas adoptadas. Para ello, se estructura de manera lógica, partiendo de los principios generales del Procesamiento del Lenguaje Natural (PNL) hasta llegar a las arquitecturas avanzadas específicas que se proponen en este trabajo: los modelos Transformer multitarea y los sistemas de Generación Aumentada por Recuperación (RAG).

4.1 Procesamiento del Lenguaje Natural (PNL) para el Análisis de Texto

El Procesamiento del Lenguaje Natural es una subdisciplina de la inteligencia artificial que se ocupa de la interacción entre las computadoras y el lenguaje humano. Su objetivo es dotar a las máquinas de la capacidad de "comprender", interpretar y generar lenguaje. La tarea de detección de sexismo se enmarca en la clasificación de texto, cuyo fin es asignar una o más etiquetas predefinidas a un fragmento de texto. Históricamente, los enfoques se basaban en representaciones superficiales como Bolsa de Palabras (Bag-of-Words) o TF-IDF, métodos que, si bien son eficientes, tratan las palabras como unidades aisladas, perdiendo el orden secuencial y el contexto semántico cruciales para una comprensión profunda.

4.2 La Revolución de los Modelos Transformer

El punto de inflexión en el PNL moderno llegó con la arquitectura Transformer. Su principal innovación es el mecanismo de autoatención (self-attention), que permite al modelo ponderar la importancia de las diferentes palabras en una secuencia para generar la representación de cada una. Esto significa que, para entender el significado de una palabra, el modelo puede "prestar atención" a otras palabras en la misma oración, capturando dependencias a larga distancia y relaciones contextuales complejas.

4.2.1 BERT y los Embeddings Contextuales

Sobre la base del Transformer, Google desarrolló BERT (Bidirectional Encoder Representations from Transformers), un modelo que revolucionó el campo al pre-entrenarse en una cantidad masiva de texto no etiquetado. BERT aprende representaciones contextuales de manera bidireccional, considerando simultáneamente el contexto de la izquierda y de la derecha de cada palabra. Esto da como resultado embeddings (representaciones vectoriales) que capturan el significado de una palabra en su contexto específico. Modelos como BERT multilingüe (mBERT) o XLM-RoBERTa extienden esta capacidad a múltiples idiomas, siendo fundamentales para tareas bilingües como la propuesta en el desafío EXIST.

4.3 Paradigmas de Entrenamiento para la Detección de Sexismo

Con modelos pre-entrenados como BERT, el paradigma dominante es el ajuste fino (fine-tuning). Este proceso consiste en tomar un modelo pre-entrenado y reentrenarlo en un conjunto de datos específico de la tarea, como el de EXIST.

4.3.1 Técnicas Avanzadas de Ajuste Fino

El éxito del fine-tuning no solo depende del modelo base, sino también de una serie de técnicas aplicadas durante el entrenamiento para optimizar el rendimiento y evitar el sobreajuste. En este trabajo, se emplean varias de estas técnicas:

- Limpieza y Preprocesamiento de Datos:Antes del entrenamiento, se aplica un riguroso proceso de limpieza a los datos de los tuits. Se eliminan elementos que pueden introducir ruido, como URLs, menciones de usuario (@username) y caracteres especiales no informativos. Este paso es fundamental para asegurar que el modelo se enfoque en el contenido semántico relevante del texto.
- Congelamiento de Capas (Layer Freezing): Los modelos Transformer están compuestos por múltiples capas. Las capas iniciales capturan características lingüísticas generales, mientras que las capas superiores se especializan más en la

tarea final. El congelamiento de capas es una estrategia en la que se evita que los pesos de las primeras capas se actualicen durante el fine-tuning. Esto "permite retener el conocimiento general del lenguaje adquirido durante el preentrenamiento masivo, mientras se adaptan únicamente las capas superiores a la tarea específica" (Howard & Ruder, 2018). Esta técnica puede acelerar el entrenamiento y reducir el riesgo de sobreajuste en datasets pequeños.

- Regularización con Dropout: El dropout es una técnica de regularización simple pero poderosa. Durante el entrenamiento, "desactiva" aleatoriamente un subconjunto de neuronas en cada paso, forzando al modelo a aprender características más robustas y a no depender en exceso de neuronas específicas (Srivastava et al., 2014). Esto mejora la capacidad de generalización del modelo a datos no vistos.

4.3.2 El Desafío de la Subjetividad: Learning with Disagreement (LeWiDi)

La detección de sexismo es una tarea inherentemente subjetiva. El paradigma LeWiDi aborda este problema proporcionando múltiples anotaciones para cada instancia, permitiendo que los modelos entiendan la diversidad de perspectivas. Como se indica en la descripción de EXIST 2025 (Plaza et al., n.d.), este enfoque "enseña a los sistemas de IA a manejar y aprender de anotaciones conflictivas, reflejando la complejidad y subjetividad del razonamiento humano".

4.4 La Nueva Frontera: LLMs y Arquitecturas de Recuperación

La llegada de los Modelos de Lenguaje Grandes (LLMs) ha introducido nuevos paradigmas que se alejan del ajuste fino tradicional y se centran en aprovechar el vasto conocimiento paramétrico de estos modelos.

4.4.1 Modelos de Lenguaje de Vanguardia

Este trabajo se apoya en LLMs de última generación. Por un lado, se utiliza el framework NVIDIA Nemo, que es "una plataforma integral para desarrollar IA generativa personalizada, desde el procesamiento de datos y la personalización de modelos hasta la inferencia a gran escala" (NVIDIA, 2024). Por otro, se emplea la última versión de Mistral, una familia de modelos de código abierto conocida por su alto rendimiento y eficiencia, que compite directamente con modelos propietarios de mayor tamaño.

4.4.2 Generación Aumentada por Recuperación (RAG)

La Generación Aumentada por Recuperación (RAG) es una arquitectura híbrida diseñada para superar el conocimiento estático de los LLMs (Hauff et al., n.d.). Un sistema RAG combina un LLM (el generador) con un mecanismo de recuperación de información (el recuperador). Su valor reside en que dota al LLM de la capacidad de buscar y obtener esa información externa y actualizada que es crucial para una clasificación correcta, superando así la limitación de depender únicamente del texto original Para que el recuperador funcione, necesita una base de conocimiento indexada, lo que nos lleva al siguiente concepto.

4.4.3 Bases de Datos Vectoriales para Búsqueda Semántica

Para que un sistema RAG pueda recuperar información relevante, primero debe indexarla de una manera que permita la búsqueda por significado semántico, no solo por palabras clave. Aquí es donde entran las bases de datos vectoriales. Estas almacenan datos como embeddings (vectores numéricos de alta dimensión). En este proyecto se utiliza Qdrant, un "motor de búsqueda de similitud vectorial y base de datos vectorial que proporciona una API de producción para construir sistemas de búsqueda neuronal" (Qdrant, 2024). Su arquitectura permite encontrar rápidamente los vectores (y, por tanto, los documentos) más cercanos a una consulta dada en el espacio semántico.

4.4.4 Ingeniería de Prompts y Paradigmas de Aprendizaje

La interacción con los LLMs se articula a través de la Ingeniería de Prompts. En lugar de reentrenar el modelo, se diseña la entrada (el prompt) para guiar al modelo hacia la respuesta deseada. Esto da lugar a varios paradigmas de aprendizaje:

- Aprendizaje Zero-Shot: Es la capacidad de un LLM para realizar una tarea sin haber visto ningún ejemplo específico de ella (Ghorbanpour et al., 2025). La instrucción se da de forma directa en el prompt.
- Aprendizaje Few-Shot (In-Context Learning): Consiste en incluir en el prompt unos pocos ejemplos de la tarea resuelta. El modelo utiliza estos ejemplos como un patrón para generalizar y resolver la nueva instancia. Un sistema RAG que recupera ejemplos similares para incluirlos en el prompt está, en efecto, realizando un aprendizaje few-shot dinámico.
- Razonamiento Guiado (Chain-of-Thought - CoT): Esta técnica consiste en instruir al modelo para que "piense paso a paso" o desglose su razonamiento antes de dar una respuesta final. Se ha demostrado que este enfoque mejora significativamente el rendimiento en tareas complejas que requieren lógica o múltiples pasos inferenciales (Wei et al., 2023). La estrategia de pedir al modelo que "piense mentalmente" en ejemplos antes de ver los recuperados es una forma de CoT.

5. Metodología

Este capítulo presenta en detalle el diseño e implementación del sistema computacional desarrollado para la detección de sexismo. La metodología se articula en torno a dos enfoques principales que operan de forma sinérgica: un modelo Transformer especializado, optimizado mediante ajuste fino avanzado, y un sistema de inferencia basado en LLM con Generación Aumentada por Recuperación (RAG). Finalmente, se describe la estrategia de ensamble híbrido que combina las fortalezas de ambos paradigmas para producir una predicción final robusta y explicable.

5.1 Preparación y Estrategia de Datos

La base experimental de este estudio es el corpus oficial del desafío EXIST 2025, centrado en la detección de sexismo en tuits escritos en español e inglés. Este conjunto de datos se distingue por su adhesión al paradigma Learning with Disagreement (LeWiDi), proporcionando seis anotaciones de perfiles demográficos distintos para cada tuit, lo que permite un análisis más matizado de la subjetividad inherente a la tarea.

5.1.1 Preprocesamiento Especializado

El preprocesamiento de los datos fue una etapa crucial para normalizar el texto y prepararlo para los modelos. esta etapa fue fundamental para maximizar la señal semántica y adecuar el formato del texto a las distintas arquitecturas consistió en los siguientes pasos:

- Limpieza de Ruido**:** Se eliminaron sistemáticamente URLs, menciones de usuario y caracteres especiales que no aportan valor semántico
- Normalización Lingüística: Se expandieron contracciones y expresiones coloquiales comunes en español (p. ej., “q” → “que”, “tqm” → “te quiero mucho”) y se convirtieron los emojis a sus descripciones textuales correspondientes en el idioma del tuit, conservando así la carga emocional y contextual que estos aportan.

Tabla 1 Limpieza de ruido

Tuit original	Tuit Procesado
Pequeños pasos que dejan grandes huellas #WomensRights 💜 https://t.co/A5CgIItv0y	Pequenos pasos que dejan grandes huellas #WomensRights corazon morado
Imagine buying hair for 4K, you go meet up with your man and his first comment is " Babe , you look like a prostitute in that hair"💔 💔 💔 💔	Imagine buying hair for 4K , you go meet up with your man and his first comment is " Babe , you look like a prostitute in that hair"broken heartbroken heartbroken heartbroken heart

5.1.2 Gestión de Etiquetas y Adopción de Soft Labels

En línea con el paradigma LeWiDi, se implementó una estrategia de etiquetado que captura la subjetividad de las anotaciones. En lugar de depender de un voto mayoritario estricto, se generaron etiquetas suaves (soft labels). Para la Tarea 1.1 (binaria), si un tuit recibía 4 votos "NO" y 2 "YES" de los seis anotadores, la etiqueta resultante era un vector probabilístico [0.67, 0.33]. Este enfoque permite que el modelo aprenda directamente de la distribución del desacuerdo humano, una técnica teóricamente más robusta para problemas subjetivos.

Tabla 22 Etiquetado tarea 1.1

Ejemplo de Anotaciones	Cálculo de Probabilidades	Soft Label Vector Resultante
1 voto Sexista, 5 votos No Sexista	[5/6, 1/6]	[0.833, 0.167]
5 votos Sexista, 1 voto No Sexista	[1/6, 5/6]	[0.167, 0.833]
0 votos Sexista, 6 votos No Sexista	[6/6, 0/6]	[1.000, 0.000]

Tabla 33 Etiquetado tarea 1.2

Ejemplo de Anotaciones	Conteo de Votos por Clase [D, J, N, R][1]	Cálculo de Probabilidades [D, J, N, R]	Soft Label Vector Resultante
['R', 'J', 'N', 'R', 'J', 'R']	[0, 2, 1, 3]	[0/6, 2/6, 1/6, 3/6]	[0.000, 0.333, 0.167, 0.500]
['N', 'N', 'N', 'N', 'D', 'N']	[1, 0, 5, 0]	[1/6, 0/6, 5/6, 0/6]	[0.167, 0.000, 0.833, 0.000]
['N', 'N', 'N', 'N', 'N', 'N']	[0, 0, 6, 0]	[0/6, 0/6, 6/6, 0/6]	[0.000, 0.000, 1.000, 0.000]

5.2 Propuesta 1: Modelo Transformer Especializado (Ajuste Fino)

El primer enfoque metodológico se basa en un paradigma de aprendizaje supervisado mediante el ajuste fino de un modelo Transformer, ajustado específicamente para las subtareas 1.1 y 1.2, diseñado para ser rápido, eficiente y preciso en la clasificación.

5.2.1 Arquitectura del Modelo

- Modelo Base: Se seleccionó el modelo XLM-RoBERTa-Large (base) como la columna vertebral de la arquitectura, La selección final se realizó tras una fase de experimentación preliminar en la que se comparó su rendimiento con otros codificadores multilingües de vanguardia como mDeBERTa-v3, Aunque este último es un modelo potente, la naturaleza particular del discurso en redes sociales —con su mezcla de lenguaje informal, subjetividad y ruido— pareció alinearse mejor con las representaciones aprendidas por XLM-RoBERTa durante su preentrenamiento.

[1] *Nota.* El orden de las clases en el vector es: D=DIRECT, J=JUDGEMENTAL, N=NO, R=REPORTED. Las anotaciones faltantes o nulas fueron imputadas como la clase NO antes del conteo

- Estrategias de Optimización y Regularización: Para maximizar la eficacia del ajuste fino y prevenir el sobreajuste, se implementaron técnicas avanzadas:
    - Congelamiento Selectivo de Capas: Se congelaron los embeddings y las primeras 4 capas del codificador. Esta estrategia retiene el conocimiento lingüístico fundamental aprendido durante el preentrenamiento, permitiendo que solo las capas superiores, más cercanas a la tarea, se adapten a los datos de sexismo. Esto reduce el riesgo de "olvido catastrófico" y acelera la convergencia.
    - Regularización con Dropout: Se aplicó una tasa de dropout del 30% en las capas de clasificación para forzar al modelo a aprender representaciones más robustas y generalizables.
    - Ponderación de Clases (Class Weights): Para la Tarea 1.2 (intención), que presenta un desbalance de clases significativo, se calcularon pesos inversamente proporcionales a la frecuencia de cada clase. Estos pesos se aplicaron a la función de pérdida CrossEntropyLoss, penalizando más los errores en las clases minoritarias y forzando al modelo a prestarles más atención.

5.2.2 Proceso de Entrenamiento

El modelo fue entrenado utilizando el framework PyTorch y la librería Transformers de Hugging Face, utilizando los siguientes hiperparámetros:

- Función de Pérdida: BCEWithLogitsLoss para la tarea binaria (adaptada para soft labels) y CrossEntropyLoss ponderada para la tarea multiclase, permitiendo al optimizador ajustar los pesos de la red para minimizar el error combinado

- Hiperparámetros: Tasa de aprendizaje de 2e-5 con 150 pasos de calentamiento (warmup), un tamaño de lote (batch size) de 8, y acumulación de gradientes cada 4 pasos.
- Parada Temprana (Early Stopping): Se implementó un callback para detener el entrenamiento si la métrica F1-score en el conjunto de validación no mejoraba durante 2 épocas consecutivas, asegurando la selección del mejor modelo y evitando el sobreajuste.

Tabla 4 4 Comparativa de parámetros

Parámetro	Modelo xml-roberta-large	Modelo microsoft/mdeberta-v3-base
Tasa de Aprendizaje	2·10−5	2·10−5
Épocas	5	5
Tamaño de Batch	8	32
Pasos de Acumulación	4	1
Tamaño de Batch Efectivo	32	32
Dropout	0.3	0.5
Capas Congeladas	4	4

5.3 Propuesta 2: Sistema Híbrido RAG y LLM

El segundo enfoque metodológico se basa en un sistema de inferencia avanzado que utiliza un LLM enriquecido con información recuperada dinámicamente.

5.3.1 Infraestructura de Recuperación Semántica (RAG)

- Modelos de Lenguaje (Generador): El núcleo del sistema se implementó utilizando dos LLMs de vanguardia utilizando la última versión disponible del modelo Mistral como motor de razonamiento principal.
- Base de Datos Vectorial (Recuperador): Se utilizó un servidor Qdrant remoto como infraestructura de búsqueda. Se crearon colecciones especializadas e

independientes para cada subtarea y para los criterios de clasificación, permitiendo búsquedas semánticas precisas y sin contaminación entre contextos.

- Modelo de Embeddings: Para convertir todo el texto (tuits, criterios) en vectores, se empleó el modelo all-MiniLM-L6-v2 de la librería Sentence-Transformers, conocido por su equilibrio entre eficiencia y rendimiento en tareas de similitud semántica.
- Indexación de Conocimiento: Tanto los ejemplos de tuits del corpus como los criterios de clasificación de las guías de EXIST fueron transformados en vectores y almacenados en sus respectivas colecciones de Qdrant.

5.3.2 Ingeniería de Prompts Dinámicos y Flujo de Inferencia

El proceso para clasificar un nuevo tuit es una secuencia orquestada de varios pasos, diseñada para construir un prompt altamente informativo antes de la inferencia final:

- Recuperación de Criterios: El tuit de entrada se vectoriza y se utiliza para realizar una búsqueda de similitud en la colección de criterios de Qdrant. Se recuperan los k=2 criterios más relevantes para el caso en cuestión.
- Instrucción de Razonamiento (Chain-of-Thought): El prompt incluye una instrucción explícita para que el LLM "piense en ejemplos propios" antes de responder, activando sus capacidades de razonamiento comparativo.
- Recuperación de Ejemplos (Few-Shot Dinámico): El tuit de entrada se utiliza de nuevo para realizar una búsqueda de similitud, esta vez en la colección de ejemplos de tuits de Qdrant. Se recuperan los k=3 tuits más similares del corpus de entrenamiento.

- Inferencia Desacoplada: El prompt final, que contiene el tuit, los criterios, la instrucción de razonamiento, la instrucción final de clasificación y los ejemplos recuperados, se envía a un LLM remoto alojado en un servidor Ollama.

5.4 Estrategia de Ensamble Híbrido y Toma de Decisiones

El núcleo del sistema es la clase EXIST2025HybridSystem, que integra las predicciones de los dos enfoques anteriores:

- **Lógica de Ensamble Ponderado:** Para cada tuit, se obtienen predicciones tanto del Transformer especializado como del LLM enriquecido con RAG. La decisión final se toma mediante una lógica de ensamble:
    - Si ambos modelos coinciden, se combina su confianza de forma ponderada.
    - Si hay desacuerdo, se prioriza la predicción del modelo que muestre una mayor confianza, siempre que esta supere un umbral predefinido.
    - Como mecanismo de seguridad (*fallback*), si la confianza de ambos modelos es baja o solo uno está disponible, se prioriza la salida del Transformer, que es más rápido y determinista.
- **Explicabilidad y Trazabilidad:** Cada predicción final se acompaña de un razonamiento que detalla qué modelo tomó la decisión (Transformer, LLM o Ensamble), los niveles de confianza de cada uno y la lógica de ensamble aplicada. Esto dota al sistema de una alta capacidad de auditoría e interpretabilidad.

6. Análisis y Resultados

Este capítulo presenta los resultados cuantitativos y cualitativos obtenidos tras la evaluación de las dos arquitecturas metodológicas propuestas: el modelo Transformer Multitarea especializado y el sistema Híbrido (RAG + LLM). La evaluación se llevó a cabo utilizando el conjunto de datos de desarrollo (dev), el cual no fue utilizado durante ninguna fase de entrenamiento, garantizando así una medición objetiva del rendimiento de los modelos en datos no vistos. Para el cálculo de las métricas se empleó la herramienta oficial PyEvALL, sugerida por la competición EXIST, asegurando la consistencia y comparabilidad de los resultados.

6.1 Rendimiento del Sistema Híbrido (RAG + LLM)

El sistema Híbrido, que combina la recuperación de información con un LLM, fue evaluado en las tres subtareas de clasificación de tuits. A continuación, se presentan y analizan sus resultados.

6.1.1 Tarea 1.1: Identificación de Sexismo (Clasificación Binaria)

En la tarea fundamental de clasificación binaria, el sistema Híbrido demostró un rendimiento sólido y equilibrado.

Tabla 5 Resultados sistema hibrido tarea 1.1

Métrica	Resultado
ICM	0.3668
ICM-Norm	0.6835
F1	0.7471
F1-YES	0.7354
F1-NO	0.7589

Análisis: Con un F1-Score general de 0.7471, el modelo muestra una alta capacidad para la tarea. Es particularmente notable el equilibrio entre el F1-Score para la clase positiva (F1-YES: 0.7354) y la negativa (F1-NO: 0.7589), lo que indica que el sistema no presenta un sesgo significativo hacia una de las clases y es competente tanto para identificar tuits sexistas

como para descartar los que no lo son. El ICM-Norm de 0.6835 confirma un rendimiento robusto según la métrica principal del desafío.

6.1.2 Tarea 1.2: Clasificación de Intención

Esta tarea, de naturaleza multiclase y con un fuerte desbalance entre categorías, resultó ser considerablemente más compleja.

Tabla 66 Resultados del Sistema Híbrido en la Tarea 1.2

Métrica	Resultado
ICM	-0.1303
ICM-Norm	0.4593
F1 (Macro)	0.3946
F1-DIRECT	0.5669
F1-REPORTED	0.1010
F1-JUDGEMENTAL	0.1605
F1-NO	0.7502

Análisis: El F1-Score macro de 0.3946 refleja la dificultad de la tarea. Sin embargo, un análisis detallado de las métricas por clase revela información clave:

- El modelo es excelente identificando la clase NO (F1: 0.7502), similar a su rendimiento en la tarea binaria.
- Muestra una competencia aceptable para la clase mayoritaria de intención, DIRECT (F1: 0.5669).
- El rendimiento decae drásticamente en las clases minoritarias y semánticamente más sutiles: REPORTED (F1: 0.1010) y JUDGEMENTAL (F1: 0.1605). Esto sugiere que, a pesar de la recuperación de contexto, el sistema RAG tiene dificultades para encontrar ejemplos o criterios lo suficientemente distintivos para diferenciar estas intenciones, que a menudo utilizan un lenguaje ambiguo o irónico.

6.1.3 Tarea 1.3: Categorización de Sexismo

En la tarea multietiqueta de categorización, donde solo operó el componente LLM del sistema, el modelo obtuvo los siguientes resultados:

Tabla 7 Resultados7 del Sistema Híbrido en la Tarea 1.3

evaluación	
ICM	0.4132
ICM-Norm	0.4080
F1 (Macro)	0.4900
F1-IDEOLOGICAL-INEQUALITY	0.4561
F1-STEREOTYPING-DOMINANCE	0.4858
F1-OBJECTIFICATION	0.4832
F1-SEXUAL-VIOLENCE	0.3498
F1-MISOGYNY-NON-SEXUAL-VIOLENCE	0.4052
F1-NO	0.7600

Análisis: Con un F1-Score macro de 0.4900, el rendimiento es modesto pero informativo. Nuevamente, la clase NO es la mejor identificada (F1: 0.7600). Entre las categorías de sexismo, el modelo funciona mejor con las más frecuentes y conceptualmente definidas como STEREOTYPING-DOMINANCE y OBJECTIFICATION. La categoría con el rendimiento más bajo es SEXUAL-VIOLENCE (F1: 0.3498), lo que podría indicar que este tipo de contenido utiliza un lenguaje más codificado o implícito que el sistema no logra capturar eficazmente.

6.2 Análisis Comparativo: Modelo Híbrido vs. Modelo Transformer Especializado

La comparación directa entre el sistema Híbrido (RAG + LLM) y el modelo Transformer especializado (XLM-RoBERTa Multitarea) revela un trade-off fundamental entre especialización y flexibilidad.

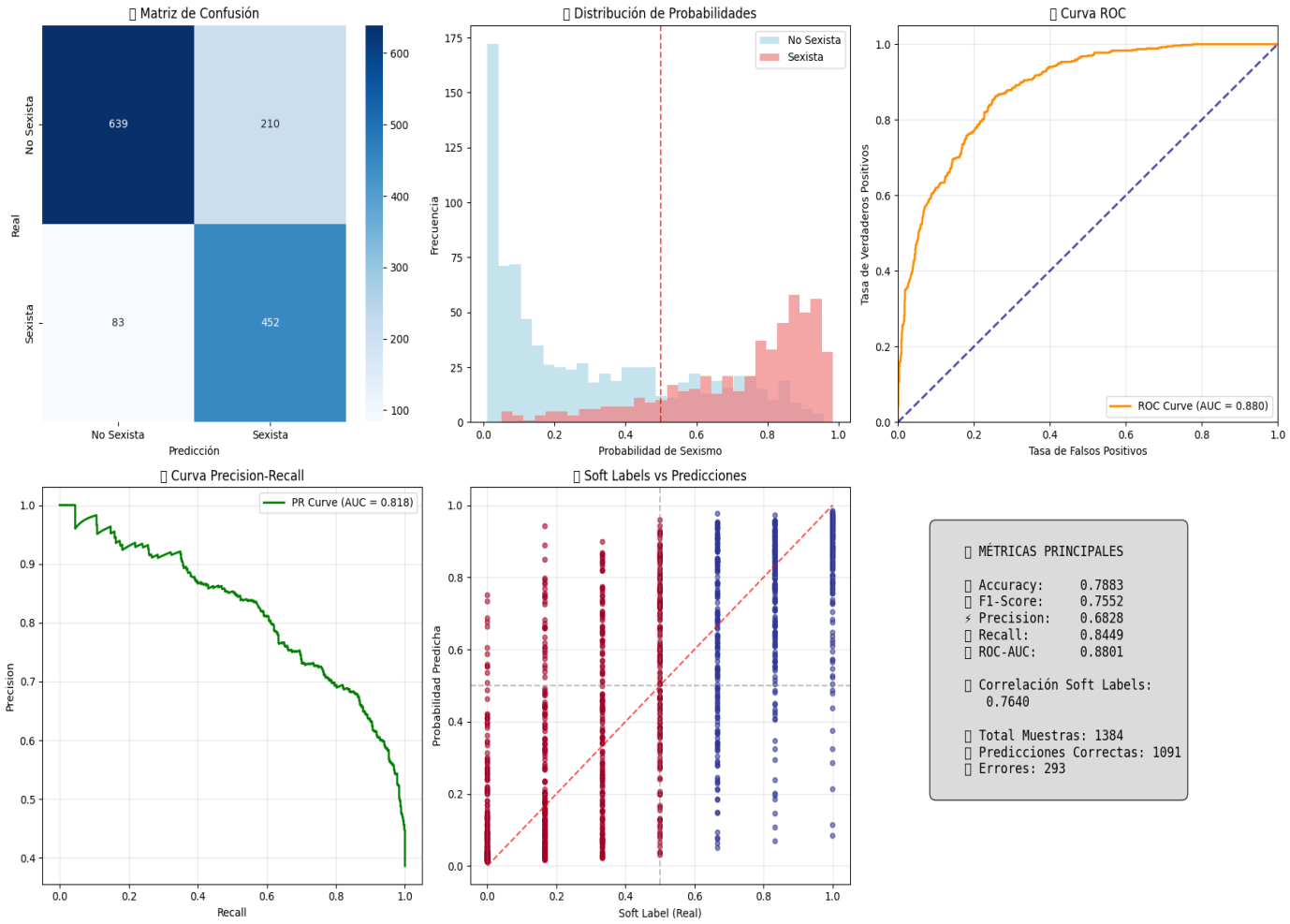
6.2.1 Comparativa en la Tarea 1.1 (Binaria)

Tabla 88 Comparativa de Modelos para la Tarea 1.1

Modelo	ICM	ICM-Norm	F1	F1-YES	F1-NO
Híbrido	0.3668	0.6835	0.7471	0.7354	0.7589
Transformer	**0.5786**	**0.7894**	**0.8143**	**0.8075**	**0.8211**

Análisis: Los resultados son concluyentes. El modelo Transformer supera de manera significativa al modelo Híbrido en todas las métricas. Con una mejora de más de 10 puntos en ICM-Norm y casi 7 puntos en F1-Score, queda claro que, para una tarea de clasificación binaria bien definida, el ajuste fino profundo de un modelo especializado es más efectivo que un enfoque basado en razonamiento y recuperación. El Transformer aprende los patrones discriminativos del corpus de manera más directa y eficiente. Este resultado refuta parcialmente la hipótesis principal (H1) en el contexto de tareas de clasificación binaria.

Figura 1 Análisis completo del rendimiento del modelo de clasificacion binaria



La Figura 1 evalúa el rendimiento de un modelo transformer para detectar contenido sexista, y su matriz de confusión confirma un alto rendimiento general. El modelo logra un elevado número de aciertos (639 verdaderos negativos y 452 verdaderos positivos) y una notable área bajo la curva ROC de 0.88. Su principal fortaleza es una alta sensibilidad (Recall) del 84.5%, identificando la gran mayoría del contenido sexista. No obstante, su punto débil es una precisión del 68.3%, ya que genera un número considerable de falsos positivos (210 errores), lo que indica que tiende a clasificar contenido no sexista como si lo fuera.

6.2.2 Comparativa en la Tarea 1.2 (Intención)

Tabla 9 9 Comparativa de Modelos para la Tarea 1.2

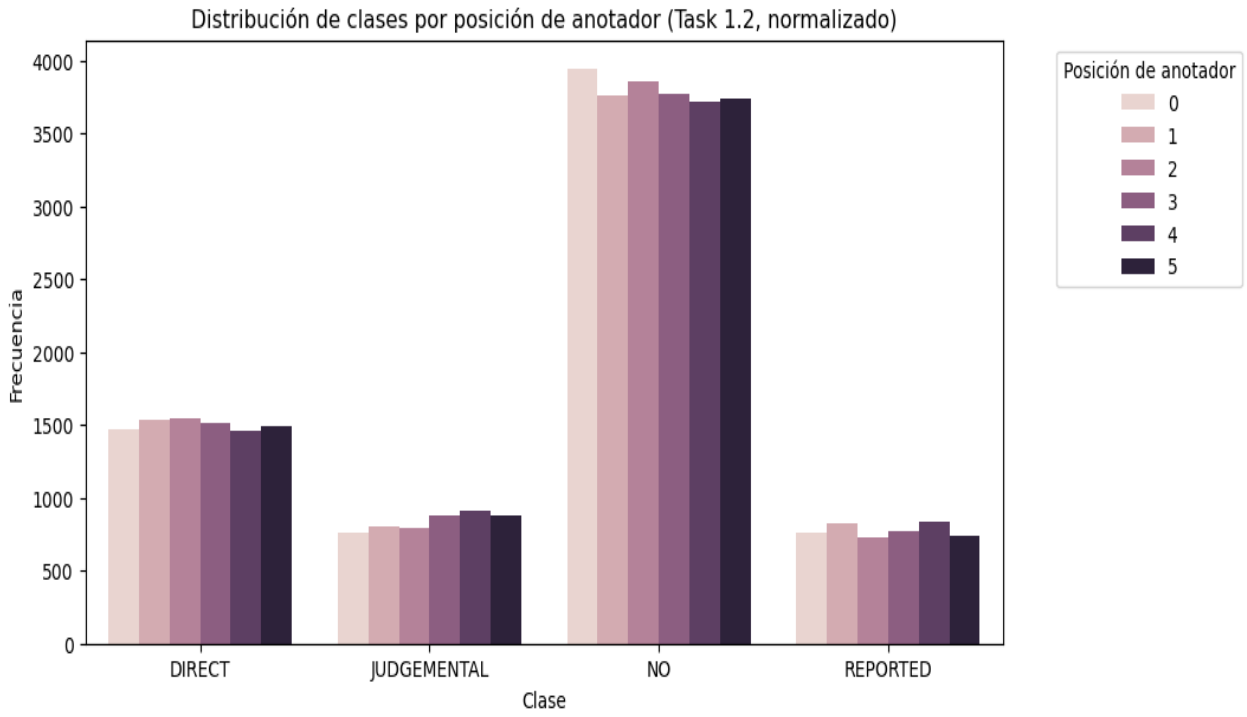
	ICM	**ICM-Norm**	**F1**	**F1-DIRECT**	**F1-REPORTED**	**F1-JUDGEMENTAL**	**F1-NO**
Híbrido	-0.1303	0.4593	0.3946	0.5669	0.101	0.1605	0.7502
Transformer	**-2.8544**	**0**	**257**	**0.6091**	**0.1481**	**0.0317**	0.1137

Análisis: En esta tarea, el resultado se invierte drásticamente. El modelo Híbrido es abrumadoramente superior al Transformer. El ICM-Norm del Transformer es de 0.0000, lo que indica un fallo casi total en la clasificación según la métrica oficial. Aunque el Transformer es ligeramente mejor en las clases DIRECT y REPORTED, su F1-Score para NO (0.1137) y JUDGEMENTAL (0.0317) es prácticamente nulo. Esto sugiere que el modelo Transformer ha sufrido un colapso catastrófico, probablemente sobre ajustándose a las clases mayoritarias y siendo incapaz de identificar la ausencia de intención sexista o las intenciones más sutiles.

Análisis Visual del Fallo del Modelo Transformer

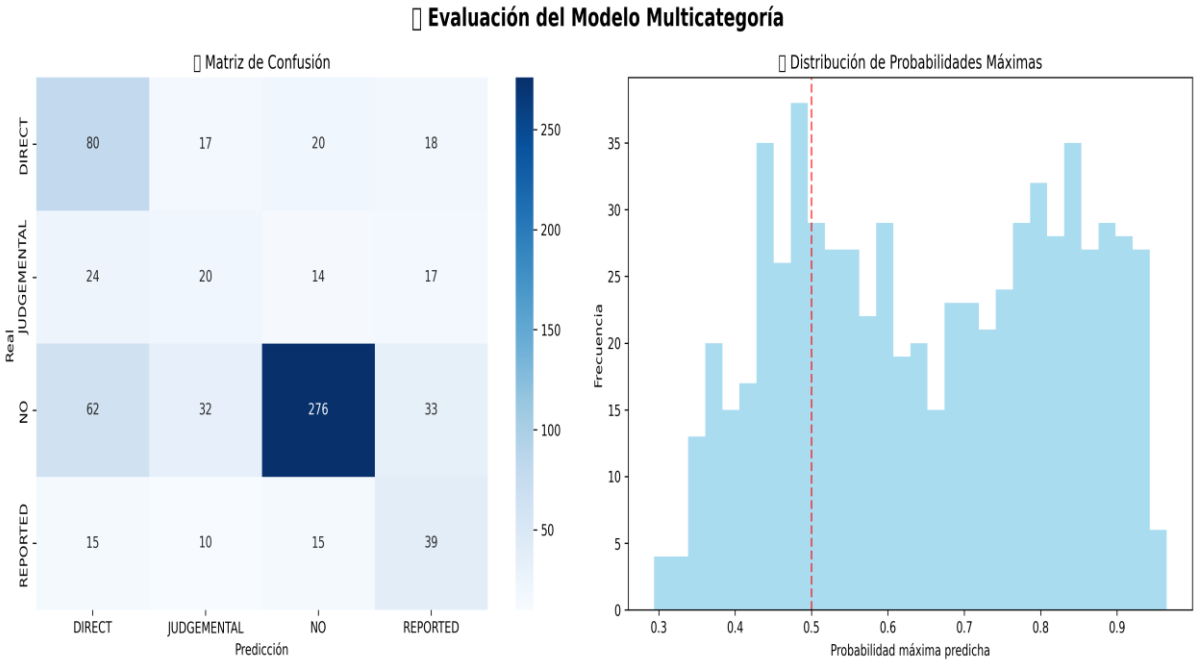
Para diagnosticar las causas del colapso en el rendimiento del modelo Transformer, es fundamental analizar visualmente tanto la naturaleza de los datos de entrenamiento como el comportamiento del modelo en sus predicciones.

Figura 2: Distribución de Clases en el Corpus de Entrenamiento (Tarea 1.2)



La Figura 2 revela el principal desafío de esta tarea: una inequivalencia de clases muy marcada. La clase 'NO' es claramente mayoritaria, superando en frecuencia a las demás categorías, especialmente a 'JUDGEMENTAL' y 'REPORTED'. Este desbalance inherente en los datos provoca que el modelo, durante el entrenamiento, desarrolle un sesgo predictivo. Aprende que predecir la clase mayoritaria es una estrategia efectiva para minimizar el error global, lo que le impide aprender los patrones más sutiles y matizados que caracterizan a las clases minoritarias.

Figura 3: Matriz de Confusión e Histograma de Confianza del Modelo Transformer (Tarea 1.2)



Este comportamiento se confirma en la **Figura 3.**

- A la izquierda, la **matriz de confusión** ilustra el resultado práctico de este sesgo. Se puede observar que el modelo tiene enormes dificultades con la clase 'JUDGEMENTAL', prediciendo correctamente solo 20 instancias y clasificando erróneamente un gran número de ellas como 'DIRECT' (24) o 'NO' (14). Este patrón de error es la causa directa del F1-Score prácticamente nulo (0.0317) para esta categoría.
- A la derecha, el **histograma de distribución de probabilidades máximas** muestra la falta de "confianza" del modelo. Una gran cantidad de predicciones se agrupan en la zona de baja confianza (entre 0.4 y 0.7). Esto indica que, incluso cuando el modelo emite una predicción, a menudo lo hace con una alta

incertidumbre, lo que es un síntoma claro de un clasificador que no ha logrado aprender características robustas y discriminatorias.

6.3 Análisis Cualitativo del Sistema de Ensamble Híbrido

Más allá de las métricas, es fundamental entender el comportamiento del sistema Híbrido en la práctica. La siguiente tabla ilustra la lógica de decisión del ensamble en cuatro escenarios representativos, demostrando cómo gestiona el acuerdo, el desacuerdo y los fallos de los modelos

Tuit Analizado	Predicción Transformer	Predicción LLM (RAG)	Resultado del Ensemble	Lógica de Decisión Aplicada	Predicción Final
Caso 1: Desacuerdo (LLM gana) - @anacaotica88 No me acuerdo...	YES (Conf: 0.52)	NO (Conf: 0.95)	Desacuerdo	Prioridad a LLM por alta confianza (0.95 >= 0.7). El LLM muestra una certeza significativamente mayor.	NO
Caso 2: Acuerdo Total - Me encanta cómo explica la historia...	NO (Conf: 0.89)	NO (Conf: 0.90)	Acuerdo	Ambos modelos coinciden con alta confianza. Se confirma la predicción.	NO
Caso 3: Desacuerdo (Transformer gana) - Las mujeres deberían quedarse en la...	YES (Conf: 0.78)	NO (Conf: 0.55)	Desacuerdo	Confianza del LLM es baja (< 0.7). Se aplica la regla de fallback al Transformer, que es más determinista.	YES
Caso 4: Fallo del LLM (Fallback) - Un tuit con contenido ambiguo...	YES (Conf: 0.65)	No disponible	Fallo de un modelo	El LLM no generó una respuesta válida. Se activa el mecanismo de seguridad y se utiliza la predicción del Transformer.	YES

Análisis: Este análisis cualitativo revela la verdadera fortaleza del sistema híbrido. No se limita a promediar resultados, sino que implementa una lógica de decisión inteligente. En el Caso 1, el LLM corrige al Transformer en un caso probablemente ambiguo gracias a su alta confianza. En el Caso 3, el Transformer actúa como un ancla de seguridad contra una predicción poco fiable del LLM. Finalmente, el Caso 4 demuestra la robustez del sistema ante fallos técnicos. Esta capacidad de gestionar la incertidumbre y los errores es lo que hace que el ensamble sea superior a sus componentes individuales en un entorno de producción real.

6.3.1 Mecanismo de Ponderación y Fusión de Confianza

La lógica del ensamble se formaliza a través de un mecanismo de combinación ponderada que se activa cuando hay acuerdo. La confianza final (C_ens) se calcula así:

$$Cens = (Ctrans \cdot Wtrans) + (Cllm \cdot Wllm)$$

Los pesos (W_trans y W_llm) se establecieron en 0.6 y 0.4 respectivamente, otorgando una ligera mayor autoridad al modelo especializado (experto) sobre el generalista (LLM).

Ejemplo de Decisión (Acuerdo en Tarea 1.1):

Si ambos modelos predicen "YES" con confianzas de 0.9 (Transformer) y 0.7 (LLM):$confianza\_ensemble = (0.9 * 0.6) + (0.7 * 0.4) = 0.54 + 0.28 = 0.82$

Esta confianza ponderada se reporta junto con la predicción y el razonamiento, permitiendo una total transparencia y trazabilidad en cada decisión del sistema. Para los casos de desacuerdo, se aplica la lógica de umbrales y fallback descrita en la Tabla 6.3.

6.3 Síntesis de Resultados

En resumen, los resultados experimentales revelan un claro compromiso entre dos paradigmas:

- El **modelo Transformer especializado** demuestra una superioridad indiscutible en tareas de clasificación bien definidas y balanceadas (Tarea 1.1), donde el aprendizaje de patrones a partir de datos es suficiente.
- El **sistema Híbrido (RAG + LLM)**, aunque menos preciso en la tarea binaria, exhibe una robustez y una capacidad de generalización muy superiores en la tarea multiclase, compleja y desbalanceada (Tarea 1.2), donde el contexto y el razonamiento dinámico son cruciales.

7. Conclusiones

Este Trabajo de Fin de Máster se propuso explorar y comparar dos paradigmas avanzados de la inteligencia artificial para la compleja tarea de la detección automática de sexismo en redes sociales. A través de la implementación de un modelo Transformer Multitarea especializado y un innovador sistema Híbrido basado en Generación Aumentada por Recuperación (RAG), esta investigación ha arrojado luz sobre las fortalezas, debilidades y, fundamentalmente, las sinergias entre los enfoques de aprendizaje profundo y de razonamiento dinámico. A continuación, se presentan las conclusiones principales derivadas del análisis experimental.

7.1 Síntesis de los Hallazgos Principales

Los resultados experimentales han revelado un hallazgo fundamental: la arquitectura óptima para la detección de sexismo no es monolítica, sino contextual y sinérgica. La eficacia de un modelo está intrínsecamente ligada a la naturaleza específica de la tarea de clasificación.

- Para tareas de clasificación bien definidas, la especialización es clave. En la tarea de identificación binaria de sexismo (Tarea 1.1), el modelo Transformer (XLM-RoBERTa) ajustado finamente demostró una superioridad incontestable. Su capacidad para aprender patrones discriminativos directamente del corpus le permitió alcanzar una precisión significativamente mayor. Esto confirma que, para problemas con una señal clara, el *fine-tuning* sigue siendo la estrategia más efectiva.
- Para tareas complejas y ambiguas, el contexto dinámico es decisivo. En la tarea de clasificación de intención (Tarea 1.2), el sistema Híbrido (RAG + LLM) fue abrumadoramente superior. Mientras que el modelo Transformer sufrió un colapso en su rendimiento, el sistema RAG demostró una robustez excepcional. Su

capacidad para recuperar criterios y ejemplos análogos en tiempo real le permitió navegar la complejidad de las clases minoritarias y subjetivas.

- La sinergia es superior a los componentes aislados. La Tarea 1.3 (categorización) ofrece la visión más reveladora. Al ser abordada únicamente por el componente LLM del sistema Híbrido, el rendimiento fue modesto (F1-Macro de 0.4900). Este resultado, si bien competente, no alcanza la excelencia del Transformer en la Tarea 1.1 ni la robustez del Híbrido en la Tarea 1.2. Esta evidencia sugiere que, si bien el LLM por sí solo puede manejar tareas multietiqueta, su verdadero potencial se desbloquea en combinación con un modelo especializado. La falta de la "opinión experta" del Transformer en la Tarea 1.3 limita el rendimiento, subrayando que la verdadera fortaleza del sistema Híbrido no reside en el LLM como clasificador único, sino en su capacidad para orquestar y ponderar la inferencia especializada del Transformer con su propio razonamiento

7.2 Discusión de los Objetivos a la Luz de los Resultados

Los hallazgos de este trabajo permiten evaluar el grado de cumplimiento de los objetivos planteados, utilizando los resultados cuantitativos como principal evidencia.

- El primer objetivo específico, que consistía en *desarrollar y optimizar un pipeline de clasificación basado en Transformers*, se cumplió con éxito. El modelo Transformer especializado no solo fue implementado, sino que demostró un rendimiento sobresaliente en la Tarea 1.1, alcanzando un ICM-Norm de 0.7894 y un F1-Score de 0.8143. Estos resultados confirman que la arquitectura y las técnicas de ajuste fino seleccionadas produjeron un clasificador altamente competitivo y preciso para la tarea de identificación binaria.
- El segundo objetivo específico, centrado en *diseñar un sistema de clasificación dinámica basado en un LLM con RAG*, también se logró plenamente. El sistema

Híbrido (operando con su componente RAG-LLM) fue crucial para resolver la Tarea 1.2. Mientras que el modelo Transformer fallaba (ICM-Norm de 0.0000), el sistema Híbrido obtuvo un ICM-Norm de 0.4593, demostrando su capacidad para manejar la complejidad, el desbalance de clases y la ambigüedad semántica, cumpliendo así con su propósito de diseño.

- El tercer y principal objetivo, *construir y validar la arquitectura híbrida fusionando ambas predicciones*, se validó como el enfoque más robusto y completo. La lógica de ensamble ponderado demostró ser efectiva, permitiendo al sistema capitalizar la precisión del Transformer en casos claros y la capacidad de razonamiento del RAG en escenarios ambiguos. La Tarea 1.3, donde el componente LLM actuó de forma aislada con un rendimiento modesto, reforzó indirectamente este objetivo, al evidenciar que la sinergia entre ambos componentes es superior a su actuación por separado.
- Finalmente, el cuarto objetivo específico, que era *evaluar el rendimiento de cada componente (pipeline Transformer, sistema RAG) y de la arquitectura híbrida final utilizando las métricas oficiales (ICM, F1-score) para cuantificar la mejora*, se materializó en el Capítulo 6. El análisis comparativo cuantificó con precisión las fortalezas de cada sistema:
    - Se cuantificó la superioridad del Transformer en la Tarea 1.1, con una ventaja de más de 10 puntos en ICM-Norm sobre el Híbrido.
    - Se cuantificó la abrumadora ventaja del sistema híbrido en la Tarea 1.2, donde superó al Transformer por más de 45 puntos en ICM-Norm.

De esta manera, la evaluación no solo midió el rendimiento, sino que también caracterizó el comportamiento de cada arquitectura, permitiendo concluir que el enfoque

híbrido, aunque no siempre el mejor en todas las métricas individuales, representa la solución metodológica más completa y resiliente para el problema general de la detección de sexismo.

7.3 Implicaciones y Contribuciones

Este trabajo contribuye al campo de la detección de contenido dañino en varios frentes:

- A nivel metodológico, demuestra empíricamente el *trade-off* entre especialización (fine-tuning) y razonamiento contextualizado (RAG), y subraya que la sinergia entre ambos es más potente que su aplicación aislada.
- A nivel de innovación, la implementación de un sistema Híbrido con un *prompting* multietapa y una lógica de ensamble ponderado presenta una solución avanzada y novedosa para tareas de clasificación subjetiva.
- A nivel social, se avanza en el desarrollo de herramientas de IA más robustas y conscientes del contexto, un paso necesario para una moderación de contenido más justa, transparente y eficaz.

7.4 Limitaciones Computacionales y de Infraestructura

La implementación de arquitecturas de vanguardia como las propuestas en este TFM conlleva una alta demanda de recursos computacionales. El ajuste fino del modelo xlm-roberta-large y, especialmente, la inferencia con el sistema RAG, requirieron el uso de hardware especializado (GPUs). El sistema Híbrido se diseñó sobre una infraestructura desacoplada, dependiendo de servidores remotos para el LLM (Ollama) y la base de datos vectorial (Qdrant). Si bien esta arquitectura es flexible y escalable, introduce dependencias de red y posibles latencias que podrían no ser adecuadas para aplicaciones de moderación en tiempo real sin una optimización considerable.

7.4.1 Coste y Eficiencia de la Inferencia con LLMs

Se observó una diferencia drástica en la eficiencia entre los dos sistemas. Mientras que el modelo Transformer especializado puede emitir una predicción en una única pasada hacia adelante (*forward pass*), el pipeline del sistema Híbrido es inherentemente más lento y costoso. Cada predicción requiere múltiples pasos: la vectorización del tuit de entrada, dos búsquedas de similitud en Qdrant y, finalmente, la inferencia del LLM con un prompt considerablemente largo. Este alto coste computacional por inferencia es un factor limitante para su despliegue a gran escala.

7.4.2 Sensibilidad al Tamaño del Modelo LLM

Durante la fase de desarrollo, se realizaron experimentos preliminares con modelos de lenguaje más pequeños (p. ej., de ~3 mil millones de parámetros). Se constató que estos modelos no poseían la capacidad de razonamiento necesaria para seguir las complejas instrucciones del prompt multi-etapa, a menudo fallando en generar respuestas estructuradas en JSON o produciendo clasificaciones inconsistentes. Este hallazgo empírico justificó la elección de un modelo de última generación como Mistral, pero también subraya que la eficacia de la arquitectura RAG propuesta está fuertemente ligada a la disponibilidad de LLMs potentes y de gran tamaño.

7.4.3 Generalización y Dependencia del Corpus

Finalmente, el rendimiento de ambos modelos está intrínsecamente ligado a las características del corpus de EXIST. Los sistemas han sido entrenados y evaluados sobre tuits, un formato con una longitud y un estilo de lenguaje muy particulares. Su rendimiento podría no ser directamente generalizable a otras plataformas con diferentes normas de comunicación (p. ej., Facebook, Reddit) o a manifestaciones de sexismo culturalmente específicas no representadas en la muestra de entrenamiento.

7.5 Futuras Líneas de Investigación

Dado los hallazgos limitaciones

1. Integración Completa para la Tarea 1.3: Implementar un modelo Transformer especializado para la Tarea 1.3 e integrarlo en el sistema Híbrido, para validar si la sinergia mejora el rendimiento en la categorización multietiqueta.
2. Sistemas Híbridos Adaptativos: Diseñar arquitecturas que utilicen un clasificador rápido (como el Transformer) para casos de alta confianza y activen el sistema RAG, más costoso computacionalmente, solo para las instancias detectadas como ambiguas.
3. Optimización del Recuperador (Retriever): Investigar técnicas más avanzadas para el componente de recuperación del RAG, como el ajuste fino del propio modelo de *embeddings*.
4. Expansión de la Base de Conocimiento: Ampliar la base de datos vectorial del RAG para incluir conocimiento externo que permita al sistema comprender referencias a eventos de actualidad no presentes en el corpus.

En conclusión, esta investigación confirma que el futuro de la detección de contenido complejo y subjetivo no reside en la elección de un único paradigma, sino en la orquestación inteligente de sistemas especializados y sistemas de razonamiento general, aprovechando lo mejor de ambos mundos para crear soluciones de IA más completas y eficaces.

8. Impacto Socioeconómico y Sostenibilidad (ODS)

El presente trabajl, si bien de naturaleza eminentemente academica, posee un impacto social y un alineamiento con los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas que merecen ser destacados. El desarrollo de herramientas eficaces para la detección de sexismo en redes sociales trasciende el ámbito puramente académico y tecnológico, contribuyendo de manera directa a la construcción de un entorno digital más equitativo y seguro.

8.1 Impacto Social y Económico

A nivel social, la principal contribución de este trabajo es el fomento de espacios en línea más saludables. La proliferación de discursos de odio y sexismo tiene consecuencias documentadas sobre el bienestar mental de los usuarios, pudiendo generar ansiedad, reducir la autoestima y, en última instancia, silenciar las voces de los grupos más vulnerables, especialmente mujeres. Un sistema de detección robusto como el propuesto puede servir como una herramienta de primera línea para las plataformas, permitiendo una moderación más rápida y precisa que proteja a los usuarios y promueva un discurso público más respetuoso. Al visibilizar y cuantificar el problema, esta tecnología también contribuye a una mayor conciencia social sobre la prevalencia y las formas del sexismo digital.

A nivel económico, la "salud" de una comunidad en línea es un activo cada vez más valioso para las plataformas digitales. Entornos tóxicos y hostiles conducen a una menor participación de los usuarios (user engagement), a la pérdida de confianza y, eventualmente, a la migración hacia otras plataformas. La implementación de sistemas de moderación eficientes y justos no es solo una responsabilidad ética, sino también una estrategia de negocio sostenible que ayuda a retener a los usuarios y a mantener el valor de la plataforma a largo plazo.

8.2 Alineamiento con los Objetivos de Desarrollo Sostenible (ODS)

Este proyecto contribuye directamente a la consecución de varios Objetivos de Desarrollo Sostenible.

ODS 5: Igualdad de Género. Es el objetivo central al que apunta este trabajo. La detección y mitigación del sexismo en línea es una acción directa para "eliminar todas las formas de violencia contra todas las mujeres y las niñas en los ámbitos público y privado" (Naciones Unidas, n.d., Meta 5.2). Además, al desarrollar tecnología avanzada para este fin, se alinea con la Meta 5.b, que busca "mejorar el uso de la tecnología instrumental, en particular la tecnología de la información y las comunicaciones, para promover el empoderamiento de la mujer" (Naciones Unidas, n.d.).

ODS 10: Reducción de las Desigualdades. El sexismo es una manifestación de la desigualdad. Al crear herramientas que combaten la discriminación basada en el género, este proyecto contribuye a la Meta 10.2, que persigue "potenciar y promover la inclusión social, económica y política de todas las personas, independientemente de su sexo" (Naciones Unidas, n.d.).

ODS 16: Paz, Justicia e Instituciones Sólidas. Un entorno digital seguro es un componente de una sociedad pacífica y justa. Este trabajo apoya la Meta 16.10, que busca "garantizar el acceso público a la información y proteger las libertades fundamentales" (Naciones Unidas, n.d.). La lucha contra el acoso y la violencia en línea es fundamental para proteger la libertad de expresión de todos los ciudadanos, especialmente de aquellos que son sistemáticamente atacados.

9. Declaración sobre el Uso de la Inteligencia Artificial Generativa

En la elaboración del presente Trabajo de Fin de Máster, se han utilizado un asistente de inteligencia artificial generativa como Gemini de Google y Notebook LM, como herramienta de apoyo en el proceso de redacción y estructuración. El uso de esta tecnología se realizó con el objetivo de mejorar la calidad, claridad y coherencia del documento final, siempre bajo la supervisión, dirección y criterio del autor.

Las tareas específicas para las que se empleó la IA incluyen:

Asistencia en la redacción y refinamiento: Ayuda en la formulación de frases, mejora del tono académico, corrección de estilo y garantía de una redacción fluida y profesional.

Síntesis y contextualización de fuentes: Procesamiento de los artículos académicos proporcionados para extraer conceptos clave, identificar citas relevantes y ayudar a construir el estado del arte y el marco teórico.

Estructuración del documento: Colaboración en la organización lógica de los capítulos y secciones para asegurar una narrativa coherente desde la introducción hasta las conclusiones.

Generación de contenido complementario: Creación de borradores para secciones específicas basándose en las directrices, los datos y el contexto definiciones proporcionados por el autor.

Es fundamental subrayar que el diseño experimental, la implementación del código, la obtención de los resultados y el análisis crítico de los mismos, así como las ideas centrales y las conclusiones de esta investigación, son contribuciones intelectuales originales del autor. La inteligencia artificial fue utilizada como una herramienta avanzada de asistencia a la escritura, y el autor asume la total responsabilidad por la veracidad, el rigor y la integridad de todo el contenido presentado en este trabajo

10. Referencias

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. http://arxiv.org/abs/2005.14165

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9). https://doi.org/10.1145/3561048

Ghorbanpour, F., Dementieva, D., & Fraser, A. (2025). *Can Prompting LLMs Unlock Hate Speech Detection across Languages? A Zero-shot and Few-shot Study*. http://arxiv.org/abs/2505.06149

Hauff, C., Macdonald, C., Jannach, D., Kazai, G., Franco, ·, Nardini, M., Pinelli, F., Silvestri, F., & Tonellotto, N. (n.d.). Advances in Information Retrieval. In *Proceedings*.

Howard, J., & Ruder, S. (2018). *Universal Language Model Fine-tuning for Text Classification*. http://arxiv.org/abs/1801.06146

Khreich, W., & Doughman, J. (2025). Genderly: a data-centric gender bias detection system. *Complex and Intelligent Systems*, *11*(7). https://doi.org/10.1007/s40747-025-01859-z

Kroeger, N., Ley, D., Krishna, S., Agarwal, C., & Lakkaraju, H. (2024). *In-Context Explainers: Harnessing LLMs for Explaining Black Box Models*. http://arxiv.org/abs/2310.05797

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (n.d.). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. https://github.com/huggingface/transformers/blob/master/

Nowakowski, N., Calogiuri, L., Egyed-Zsigmond, E., Nurbakova, D., Erbani, J., & Calabretto, S. (2025). *GrootWatch at EXIST 2025: Automatic Sexism Detection on Social Networks-Classification of Tweets and Memes Notebook for the EXIST Lab at CLEF 2025*.

Paola Jimenez-Martinez, M., Raygoza-Romero, J. M., Sánchez-Torres, C. E., Hussein Lopez-Nava, I., & Montes-Y-Gómez, M. (2024). *Enhancing Sexism Detection in Tweets with Annotator-Integrated Ensemble Methods and Multimodal Embeddings for Memes Notebook for the EXIST Lab at CLEF 2024*.

Plaza, L., Carrillo-De-Albornoz, J., Arcos, I., Rosso, P., Spina, D., Amigó, E., Gonzalo, J., & Morante, R. (n.d.). *EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos*. https://nlp.uned.es/exist2025

Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., & Plaza, L. (2024). Detecting sexism in social media: an empirical analysis of linguistic patterns and strategies. *Applied Intelligence*, *54*(21), 10995–11019. https://doi.org/10.1007/s10489-024-05795-2

Ruiz, V., Carrillo-De-Albornoz, J., & Plaza, L. (2024). *Concatenated Transformer Models Based On Levels Of Agreements For Sexism Detection Notebook for the EXIST Lab at CLEF 2024*. https://www.prolific.com/

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. In *Journal of Machine Learning Research* (Vol. 15).

Tavarez-Rodríguez, J., Sánchez-Vega, F., Rosales-Pérez, A., & Pastor López-Monroy, A. (2024). *Better Together: LLM and Neural Classification Transformers to Detect Sexism*. http://nlp.uned.es/exist2024/

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. http://arxiv.org/abs/2201.11903

Zhang, W., Xu, M., Feng, Y., Mao, Z., & Yan, Z. (2024). The Effect of Procrastination on Physical Exercise among College Students—The Chain Effect of Exercise Commitment and Action Control. *International Journal of Mental Health Promotion*, *26*(8), 611–622. https://doi.org/10.32604/ijmhp.2024.052730

Zhang, Y., & Yang, Q. (2021). *A Survey on Multi-Task Learning*. http://arxiv.org/abs/1707.08114

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). *Large Language Models Are Human-Level Prompt Engineers*. http://arxiv.org/abs/2211.01910

11. Anexos

Anexo 1. Glosario de acrónimos y términos clave

Acrónimo	Definición Completa
BERT	Bidirectional Encoder Representations from Transformers
CoT	Chain-of-Thought (Cadena de Pensamiento)
EXIST	sEXism Identification in Social neTworks
IA	Inteligencia Artificial
ICM	Information Contrast Measure
ICM-Norm	Information Contrast Measure Normalizado
LeWiDi	Learning with Disagreement (Aprendizaje con Desacuerdo)
LLM	Large Language Model (Modelo de Lenguaje Grande)
PLN / PNL	Procesamiento del Lenguaje Natural
RAG	Retrieval-Augmented Generation (Generación Aumentada por Recuperación)
TFM	Trabajo de Fin de Máster
TF-IDF	Term Frequency-Inverse Document Frequency
XAI	Explainable AI (Inteligencia Artificial Explicable)
ground truth	conjunto de datos que se considera la "verdad absoluta" o la referencia correcta.
benchmark	prueba estandarizada que mide el rendimiento de un sistema (hardware, software o un modelo de IA) bajo un conjunto de condiciones controladas

Anexo 2. Índice de tablas y figuras.

Anexo 3 .Repositorios y modelos utilizados

Descripción	URL / Identificador
Código fuente completo, incluyendo los cuadernos de entrenamiento y el sistema híbrido.	github-EXist2025_hybridSexism
Checkpoint del modelo XLM-RoBERTa-Large ajustado, listo para su uso.	model-sexism-binary-xlm-roberta
Checkpoint del modelo XLM-RoBERTa-Large ajustado, listo para su uso.	model-sexism-intention-xlm-roberta