Mejoras en BERT y RoBERTa para el Análisis de Textos Críticos y Conspirativos Basados en Emociones: Un Enfoque Eficiente

Anny Álvarez Nogales[a], Paula Arias Fernández[a] and Jorge del Castillo Gómez[a,3]

[a]Universidad Politécnica de Madrid, Madrid, España

ARTICLE INFO

Keywords:
Transformers
Oppositional Thinking
Bert
RoBerta
Language Models

ABSTRACT

El análisis de textos que incorporan pensamiento de oposición, como el pensamiento crítico o conspirativo, es clave para entender las dinámicas subyacentes a ideologías extremas y procesos de radicalización. A pesar del progreso en modelos de lenguaje de gran escala (LLMs), las arquitecturas preentrenadas como BERT y RoBERTa continúan siendo herramientas esenciales para tareas de clasificación de texto, debido a su capacidad de captar matices semánticos y contextuales complejos. Este trabajo presenta un enfoque innovador para optimizar las arquitecturas base de BERT y RoBERTa, buscando replicar el rendimiento de modelos de mayor tamaño, como RoBERTa Large, pero con menores costes computacionales. Para ello, se introduce un método basado en frases auxiliares y la integración de cabezas de atención adicionales. Los resultados obtenidos destacan mejoras significativas en el rendimiento de BERT base y en nuestra versión optimizada de RoBERTa, que logran acercarse al desempeño de modelos más grandes, mostrando la efectividad de nuestro enfoque.

1. Introducción

Las ideologías conspiranoicas se han convertido en el principal foco de preocupación del ente público y académico relacionado con amenazas a la democracia y el apoyo a sistemas autocráticos, según Hogg (2021). Las narrativas conspiranoicas debilitan los sistemas democráticos porque atribuyen la responsabilidad final de la crisis a agentes ocultos que escapan del control de los gobiernos.

Korenčić et al. (2024) señalan que existe un alto riesgo en los sistemas de monitorización de redes sociales a encasillar a personas en comunidades conspiranoicas, y esto se sustenta con la teoría de identidad social. La teoría social de identidad de Tajfel (1979) establece que una parte importante de la identidad de una persona proviene de los grupos a los que pertenece. Estas personas tienden a identificar grupos a los que pertenecen y grupos de los que no forman parte. Existe una inclinación natural a favorecer al grupo al que pertenece y ser crítico o desconfiado hacia los miembros de otros grupos. Por tanto, los individuos buscan mantener una identidad social positiva dentro de un grupo, influenciando sus comportamientos y actitudes sobre miembros dentro y fuera de esos grupos.

En este sentido, Douglas, Uscinski, Sutton, Cichocka, Nefes, Ang, and Deravi (2019) describen que las personas suelen sentirse atraídas a este tipo de narrativas conspiranoicas porque satisfacen necesidades psicológicas, como lo son epistémicas (búsqueda de la certeza) y las sociales (el deseo de mantener una imagen positiva del grupo al que pertenecen). Mencionan que la atracción hacia las teorías conspirativas aumenta cuando un grupo percibe que está estigmatizado y amenazado. De tal forma, las personas no solo buscan respuestas a través de estas narrativas, sino que también refuerzan su identidad social perteneciendo a esta comunidad conspirativa, protegiéndose ante amenazas externas.

Como resultado, ser considerado un conspiranoico cuando no es así puede convertirse en una amenaza a la identidad social, según Korenčić et al. (2024). Cuando un sujeto es objetivo de esta acusación, tiende a unirse a grupos conspiranoicos para sentirse socialmente aceptado y recuperar una identidad positiva social. Por ello, es crucial la diferenciación entre el pensamiento conspiranoico y crítico para no conducir erróneamente a individuos hacia las comunidades conspirativas.

Los textos conspirativos, que a menudo están relacionados con *fake news* y desinformación, se caracterizan por un lenguaje cargado de emociones negativas. Según Newman, Pennebaker, Berry, and Richards (2003), investigaciones psicológicas han demostrado que las mentiras y la desinformación están asociadas con un lenguaje cargado de emociones negativas. Por ejemplo, Kwon, Cha, Jung, Chen, and Wang (2013) encontraron que los rumores en redes sociales presentan un sentimiento significativamente menos positivo, mientras que Rubin, Conroy, Chen, and Cornwell (2016) observaron una predominancia de términos emocionales negativos en noticias falsas, en comparación con las reales. Asimismo, Zaeem, Li, and Barber (2020) aplicaron herramientas de análisis emocional en un corpus masivo de *fake news*, descubriendo una relación significativa entre emociones negativas y este tipo de contenido, mientras que las noticias reales mostraron mayor prevalencia de sentimientos positivos.

Estas características emocionales se utilizan estratégicamente para captar la atención de los lectores y aumentar la probabilidad de compartir el contenido, como describen Zhang, Song, Chen, and Jia (2022). Las emociones juegan un papel crucial en cómo la desinformación se propaga, como destaca Horner, Galletta, Crawford, and Shirsat (2023), al

ORCID(s):
[1]Email: a.anogales@alumnos.upm.es
[2]Email: paula.ariasf@alumnos.upm.es
[3]Email: jorge.delcastillo@alumnos.upm.es

observar que los lectores tienden a confiar más en noticias acordes con sus creencias.

De manera más específica, Cosgrove and Bahr (2024) identificaron que el contenido conspirativo en redes sociales exhibe mayores tasas de emociones negativas, como ira y ansiedad, en comparación con discusiones científicas. Estas emociones no solo están presentes en el discurso conspirativo, sino que también predicen niveles más altos de interacción, destacando su papel en la propagación de estas narrativas. Recientemente, se ha demostrado que los métodos de procesamiento del lenguaje natural (NLP) que incorporan análisis de emociones pueden mejorar la detección de contenido conspirativo, según Alonso, Vilares, Gómez-Rodríguez, and Vilares (2021). Basándonos en estos hallazgos, nuestro trabajo propone integrar el análisis de emociones como un componente clave para enriquecer los modelos de detección de contenido conspirativo. En particular, buscamos mejorar el rendimiento de modelos con menos parámetros, como BERT y RoBERTa base, aspirando a que alcancen un rendimiento comparable al de sus contrapartes más grandes, sin incrementar significativamente la complejidad computacional.

Este trabajo aborda las siguientes preguntas clave:

1. **Impacto del contexto emocional:** ¿Cómo influye la incorporación del análisis emocional como un componente adicional en los modelos transformers para tareas que requieren captar alta subjetividad, como la diferenciación entre textos conspirativos y críticos? Se plantea que el enriquecimiento con información emocional podría mejorar la precisión al desambiguar matices subjetivos en estos textos.
2. **Optimización de modelos base:** ¿Es posible que modelos transformers con menos parámetros, como BERT y RoBERTa base, alcancen un rendimiento comparable al de modelos más grandes mediante el uso de técnicas como el aumento de datos y la integración del contexto emocional? La hipótesis plantea que estas estrategias pueden cerrar la brecha de desempeño sin incrementar significativamente la complejidad computacional.

2. Marco Teórico

2.1. Aumento de datos y Adición contextual

En un primer análisis de los datos, identificamos un marcado desbalanceo de clases, con una notable subrepresentación de la clase *CONSPIRACY*. Para abordar esta problemática, se implementaron técnicas de aumento de datos con el objetivo de subsanar esta deficiencia. Entre las diversas metodologías exploradas, destacamos el uso de la librería *nlpaug*, que ofrece una amplia variedad de enfoques para generar datos de manera coherente. De todas las técnicas disponibles, optamos por *Backtranslation*, un método que emplea modelos preentrenados para traducir textos a otro idioma y luego devolverlos al idioma original. En nuestro caso, seleccionamos el alemán como idioma intermedio. Esta estrategia nos permitió generar nuevos textos con pequeñas variaciones respecto al conjunto original, manteniendo sus etiquetas iniciales. En este escenario, todos los textos añadidos pertenecían a la clase *CONSPIRACY*, la menos representada. Los textos resultantes eran similares a los existentes, pero no idénticos, ayudándonos a mitigar el riesgo de sobreajuste (overfitting) mientras equilibrábamos las clases.

Dada la subjetividad inherente de la tarea, consideramos añadir información contextual a los datos para enriquecer las predicciones del modelo. Inicialmente, evaluamos la posibilidad de utilizar modelos preentrenados para extraer características textuales como emociones. Sin embargo, descartamos esta opción debido a la dependencia del sesgo de los modelos seleccionados. En lugar de ello, optamos por incorporar un enfoque más controlado: prompts contextuales. Para cada texto, agregamos al final una sección que incluía palabras clave relacionadas con valores morales o emociones, extraídas a partir de palabras clave presentes en el texto. Esta estrategia proporcionó al modelo información adicional relevante, permitiéndole realizar predicciones más informadas sin introducir el sesgo propio de un modelo externo.

2.2. Modelos Transformers

Los modelos Transformer revolucionaron el campo del procesamiento del lenguaje natural (NLP) y el aprendizaje automático al introducir mecanismos innovadores que mejoran significativamente el manejo de datos secuenciales Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin (2017). La Figura 1 ilustra la arquitectura general de estos modelos.

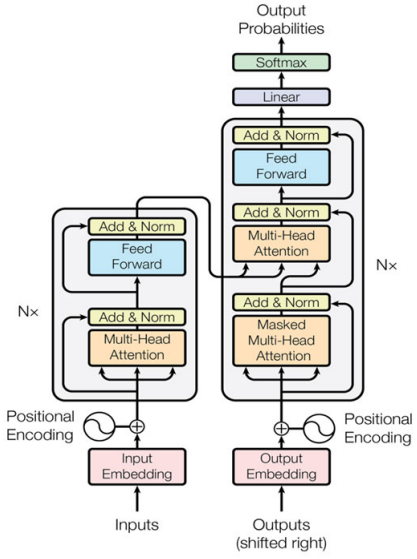

Figure 1: Arquitectura Transformers. Fuente: Vaswani et al. (2017)

Los componentes clave de los modelos Transformer son:

- **Embeddings**: Son vectores que representan las palabras de entrada, capturando sus significados semánticos y permitiendo que el modelo entienda el contexto de cada token en la secuencia.
- **Codificación Posicional**: Añadida a los embeddings para informar sobre la posición de cada token en la secuencia, ya que los Transformers no comprenden inherentemente el orden secuencial.
- **Mecanismo de Autoatención**: Permite al modelo ponderar la importancia de diferentes palabras en la secuencia, considerando su relevancia mutua.
- **Atención Multi-Cabeza**: Expande la autoatención permitiendo que el modelo atienda simultáneamente a diferentes partes de la secuencia, capturando dependencias a largo plazo.
- **Atención de Producto Escalar**: Calcula la importancia de cada palabra en relación con las demás.
- **Redes Feed-Forward**: Compuestas por capas densas que se aplican después de la atención multi-cabeza, ayudando a refinar las representaciones y capturar patrones complejos en los datos.
- **Codificador (Encoder)**: Consiste en múltiples capas que incluyen el mecanismo de autoatención y una red feed-forward, generando representaciones contextuales de las palabras en la secuencia.
- **Decodificador (Decoder)**: Similar al codificador, pero con un mecanismo adicional de atención entre el codificador y el decodificador, para generar la secuencia de salida.

2.2.1. Modelos BERT y RoBERTa

BERT (Bidirectional Encoder Representations from Transformers) es un modelo de lenguaje preentrenado desarrollado por Google en 2018 Devlin, Chang, Lee, and Toutanova (2019). Se entrena en grandes corpus de texto para predecir palabras faltantes en una secuencia, pero a diferencia de los Transformers tradicionales, BERT es entrenado de manera bidireccional, permitiendo que capture el contexto de las palabras en ambas direcciones. Sus características incluyen:

- **Enmascarado de Palabras:** Durante el entrenamiento, algunas palabras son enmascaradas, y BERT predice estas palabras basándose en el contexto circundante.
- **Clasificación Multitarea:** Además de las tareas de clasificación, BERT puede realizar inferencias y responder preguntas, utilizando un proceso de ajuste fino para diversas tareas de NLP.

RoBERTa (A Robustly Optimized BERT Pretraining Approach) es una mejora de BERT, desarrollada por Facebook AI en 2019 Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and Stoyanov (2019). Utiliza un corpus de entrenamiento más grande (160 GB frente a los 16 GB de BERT), y otros ajustes como el aumento de iteraciones de entrenamiento, la eliminación de la tarea de predicción de la siguiente oración y el uso de un enmascarado dinámico para mejorar su capacidad de generalización.

	BERT	RoBERTa
Desarrollador	Google (2018)	Meta AI (2019)
Arquitectura	Transformer Bidireccional (codificador)	Transformer Bidireccional (basado en BERT)
Tokenizador	Tokenizador WordPiece	Tokenizador SentencePiece
Corpus	Wikipedia en inglés BooksCorpus	Wikipedia en inglés BooksCorpus CC-News, OpenWebText
Parámetros	BERT base: 110M BERT large: 340M	RoBERTa base: 125M RoBERTa large: 355M

Table 1
Comparación de Modelos de Lenguaje Pre-entrenados: BERT vs RoBERTa.

La Tabla 1 resume las principales diferencias entre estos dos modelos.

3. Estado del Arte

En esta sección, se repasan los estudios más relevantes en la detección de textos conspiranoicos, desinformación y *fake news*. Se analiza la contribución de diferentes enfoques metodológicos empleados en los últimos años para abordar estos problemas, destacando el uso de modelos de lenguaje preentrenados como BERT y RoBERTa, técnicas de *data augmentation* para superar la escasez de datos etiquetados, análisis léxico para identificar signos de desinformación, y modificaciones en las arquitecturas de transformers para mejorar la precisión de las predicciones. Además, se revisan metodologías innovadoras, como la incorporación de conocimiento específico del dominio mediante el uso de frases auxiliares, que enriquecen los modelos con contexto relevante.

3.1. Modelos BERT y RoBERTa

Pavlov and Mirceva (2022) proponen la detección de *fake news* sobre COVID-19 utilizando modelos de lenguaje preentrenados como BERT y RoBERTa, aplicados a datos de *tweets* sobre la pandemia. El trabajo analiza y compara sus capacidades frente a investigaciones previas en la misma tarea, utilizando un conjunto de datos balanceado de *tweets* con noticias reales y falsas.

Los autores emplean dos modelos preentrenados: RoBERTa base, entrenado específicamente con *tweets* (*twitter-roberta-base-sentiment*), y un modelo BERT large, también preentrenado sobre datos de COVID-19 (*covid-twitter-bert*). Mientras que *twitter-roberta-base-sentiment* corresponde a una arquitectura RoBERTa base, *covid-twitter-bert* representa un modelo de mayor complejidad por su tamaño (*large*).

El estudio concluye que el *fine-tuning* tiene un impacto significativo en el desempeño de ambos modelos, mejorando considerablemente las métricas de clasificación respecto a sus configuraciones preentrenadas. Un hallazgo relevante, y

Evaluation Metric	BERT	RoBERTa
Accuracy	0.3457	0.5504
Precision	0.2971	0.6167
Recall	0.1830	0.5504
F1	0.3250	0.5743
MCC	-0.3125	0.1792

Table 2
Resultados obtenidos de los modelos BERT y RoBERTa preentrenados. Fuente: Pavlov and Mirceva (2022)

Evaluation Metric	BERT	RoBERTa
Accuracy	**0.9831**	0.9752
Precision	0.9796	0.9708
Recall	0.9883	0.9821
F1	**0.9831**	0.9752
MCC	0.9663	0.9504

Table 3
Resultados obtenidos de los modelos BERT y RoBERTa fine-tuned. Fuente: Pavlov and Mirceva (2022)

que consideramos clave para nuestra investigación, es que el modelo preentrenado RoBERTa base supera al preentrenado BERT, aunque esta ventaja varía tras el *fine-tuning*, como se puede evidenciar en la Tabla 3.1. RoBERTa sin *fine-tuning* supera con creces a BERT, pero se queda ligeramente por debajo cuando se utiliza *fine-tuning* (Tabla 3.1).

Este resultado refuerza la motivación de nuestra propuesta: optimizar modelos más compactos, como RoBERTa base, para que alcancen un desempeño comparable al de modelos *large*, independientemente de la arquitectura base utilizada.

Peskine, Papotti, and Troncy (2023) presentan una arquitectura basada en transformers para la detección de teorías conspirativas en un conjunto de datos de *tweets* relacionados con *fake news* y conspiraciones, utilizando un ensamblado de 5 modelos CT-BERT. Un CT-BERT es un modelo BERT preentrenado en un corpus extenso de datos de Twitter sobre la COVID-19, lo que lo convierte en una herramienta ideal para abordar la tarea de detección de teorías conspirativas planteada en este estudio.

Para implementar el enfoque, se entrenaron 5 modelos CT-BERT, cada uno en un *fold* diferente generado mediante validación cruzada (*cross-validation*), y se combinan mediante un ensamblado por votación mayoritaria. El *ensembling* (ensamblaje o combinación de modelos) es una técnica de aprendizaje automático que mejora el desempeño general al combinar los resultados de varios modelos individuales, aprovechando sus fortalezas y mitigando sus debilidades.

La propuesta logra una métrica de correlación de Matthews (MCC) de 0.710, destacándose como una solución efectiva para la tarea.

Albladi and Seals (2024) propone un modelo basado en RoBERTa *fine-tuned* con un *dropout* de 0.30 para abordar la tarea de diferenciación entre narrativas oposicionales, específicamente entre narrativas conspirativas y críticas, en el contexto de la competición PAN at CLEF 2024.

Table 4
Métricas de evaluación por categoría (Critical, Conspiracy). Fuente: Albladi and Seals (2024)

Category	Precision	Recall	F1 Score
CRITICAL	0.94	0.93	0.93
CONSPIRACY	0.86	0.89	0.87

El proceso de preprocesamiento incluye la tokenización, codificación (*encoding*), etiquetado (*labeling*) y la división de los datos en un 90% para entrenamiento y un 10% para prueba. El modelo fue entrenado durante 6 épocas, observándose un inicio de sobreajuste en la primera época, lo que sugiere un buen aprendizaje del modelo en los datos de entrenamiento.

En la Tabla 4, se reportan los resultados de clasificación obtenidos por el modelo, destacándose una alta precisión, recall y F1-score en la predicción de textos críticos. Estos resultados indican la eficiencia del modelo para diferenciar entre ambas categorías. Además, se alcanzó un MCC de 0.80, lo que confirma la efectividad del enfoque propuesto.

Con base en estas evidencias, justificamos la elección de BERT y RoBERTa como modelos principales en nuestra investigación, debido a su probada efectividad en el estado del arte, su adaptabilidad a datos específicos y su rol central como arquitecturas básicas en el procesamiento de lenguaje natural.

3.2. Análisis lexico

Los modelos BERT y RoBERTa han demostrado buenos resultados en tareas de clasificación binaria en la detección de teorías conspiranoicas, según lo señalado en el apartado previo. Sin embargo, aún presentan limitaciones al abordar tareas más complejas, como la diferenciación entre texto crítico y texto conspirativo.

En este contexto, Liu, Liu, Thompson, Yang, and Ananiadou (2024) proponen el primer modelo LLM de código abierto 'ConspEmoLLM', el cuál incorpora información afectiva para abordar múltiples tareas relacionadas con las teorías conspirativas. Su investigación se centra alrededor del análisis de las emociones de textos conspirativos, con el propósito de extraer información afectiva relevante y enriquecer los entrenamientos de los modelos existentes.

El análisis realizado revela que los textos relacionados con teorías conspirativas tienden a expresar emociones negativas más intensas, como ira, miedo y disgusto, mientras que los textos no relacionados exhiben mayores niveles de emociones positivas, como alegría, amor y optimismo. Por ejemplo, en los resultados del dataset COCO (Figura 2), las categorías relacionadas con teorías conspirativas muestran una intensidad más alta en emociones como ira y miedo, mientras que la tristeza se distribuye de manera similar entre todas las categorías analizadas.

El modelo ConspEmoLLM muestra resultados prometedores en comparación con otros modelos, como ConspLLM (sin enriquecimiento con emociones), ChatGPT y LLaMA2,

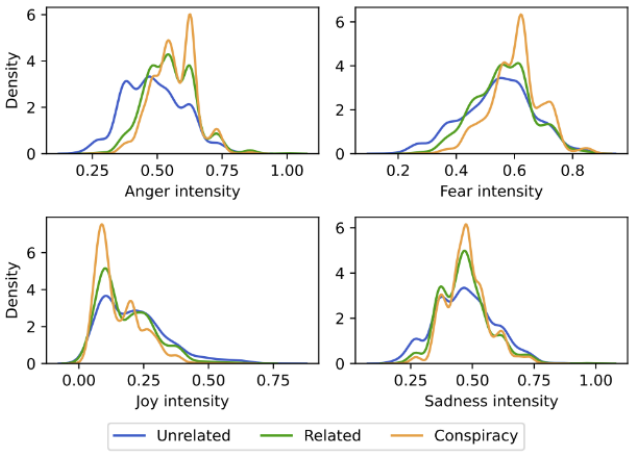

Figure 2: Intensidad de las emociones según la clase (Unrelated, Related, Conspiracy). Fuente: Liu et al. (2024)

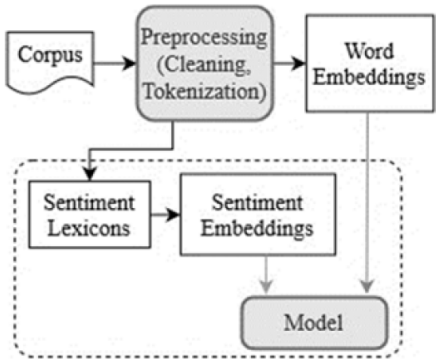

Figure 3: Diagramas de bloques del enfoque basado en léxicos de sentimientos. Fuente: Koufakou and Scott (2020)

en la tarea de determinar la relación entre un texto y una teoría conspirativa (Unrelated, Related, Conspiracy).

El uso de información emocional en ConspEmoLLM reafirma la hipótesis central del estudio de Liu et al. (2024): integrar características afectivas mejora el rendimiento de los modelos existentes en la detección y análisis de teorías conspirativas. Aunque los resultados obtenidos son mejores en comparación con los modelos previos, el rendimiento general de sus modelos aún está lejos de alcanzar un nivel óptimo.

Koufakou and Scott (2020) analizan dos enfoques para integrar léxicos con el fin de mejorar los modelos existentes en la tarea de detección de lenguaje abusivo: los léxicos de sentimientos, que incorporan información emocional a través de la integración de *embeddings* de sentimientos en el modelo de aprendizaje; y los léxicos semánticos, que transforman los embeddings de las palabras de un texto abusivo para que representen con mayor precisión las relaciones entre el lenguaje abusivo y las emociones. Los léxicos de sentimientos asignan un valor de puntuación a cada token en función de varios léxicos. Cada token se representa con un vector de dimensión l, siendo l el número de léxicos de sentimientos.

Los embeddings de las palabras se combinan con sus respectivos embeddings léxicos en la entrada del modelo, el cual aprenderá a detectar textos con lenguaje abusivo, teniendo en cuenta las emociones incrustadas (Figura 3)

Mutinda, Mwangi, and Okeyo (2023) introducen el modelo LeBERT, que combina un léxico de sentimientos, N-gramas y la representación de palabras BERT para mejorar la precisión en la clasificación de sentimientos. Este enfoque destaca por identificar secciones del texto relevantes para la información de sentimientos mediante el léxico, y aplicar embeddings preentrenados de BERT a estas secciones. En LeBERT, este léxico se aplica para seleccionar trigramas (grupos de tres palabras consecutivas) en el texto que contengan al menos una palabra que esté presente en el léxico de sentimientos. Este proceso reduce significativamente la cantidad de texto procesado, enfocándose en las secciones más significativas para el análisis de sentimientos. La arquitectura utiliza una red neuronal convolucional (CNN) para clasificar los vectores generados. Los experimentos en tres conjuntos de datos públicos (Amazon, IMDB y Yelp) muestran que LeBERT supera a modelos existentes en métricas clave como la precisión. Además, se concluye que el uso del léxico reduce la dimensionalidad de los vectores de entrada y mejora la eficiencia del modelo en tareas de clasificación binaria de sentimientos.

Las tres investigaciones mencionadas previamente sirven como un punto de partida clave para la incorporación del análisis léxico en nuestros modelos, específicamente durante la fase de implementación, con el objetivo de alcanzar las metas propuestas.

3.3. Aumento de Datos

Son varios los artículos del *state of the art* que completan los datos con diferentes técnicas. Shrirang Mhalgi and Kübler (2024) amplían los datos utilizando el corpus LOCO, el cual contiene un amplio rango de temáticas conspiranoicas, por lo cual amplía el campo temático abarcado en el dataset, que originalmente solo trata de textos del COVID-19. Por otro lado, en Angelo Maximilian Tulbure (2024) se introduce la idea de traducir textos de un idioma a otro con diferentes modelos preentrenados, como puede verse en el trabajo de Helsinki-NLP (2024), con el objetivo de aumentar los datos de entrada de inglés traduciendo los textos en español, y viceversa. En esta misma línea, se continúa el proceso en Iñaki del Campo Sánchez-Hermosilla and Camacho (2024), añadiendo además otra serie de técnicas de aumento de datos como pueden ser inserción de palabras aleatorias, sustitución de palabras o reemplazamiento por sinónimos empleando la librería nlpaug (2024). En esta línea, decidimos combinar ambas ideas con el objetivo de completar nuestros datos para mitigar el desequilibrio existente ya mencionado más arriba. Para ello hemos utilizado el aumento de datos *BackTranslationAug*, en el cual se traducen los textos a un idioma, y posteriormente se vuelve al original.

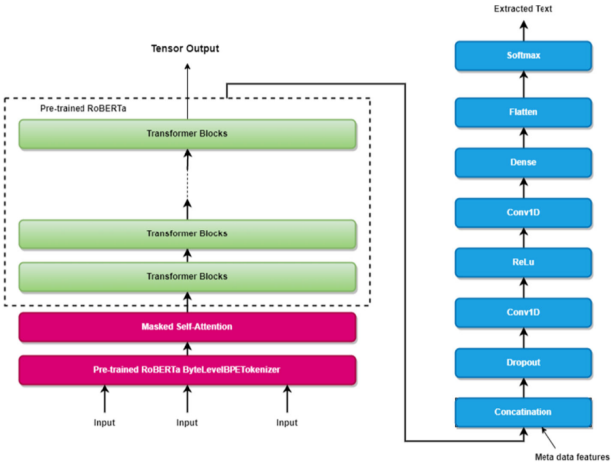

Figure 4: Modificación de la arquitectura de RoBERTa. Fuente: Cheruku et al. (2023)

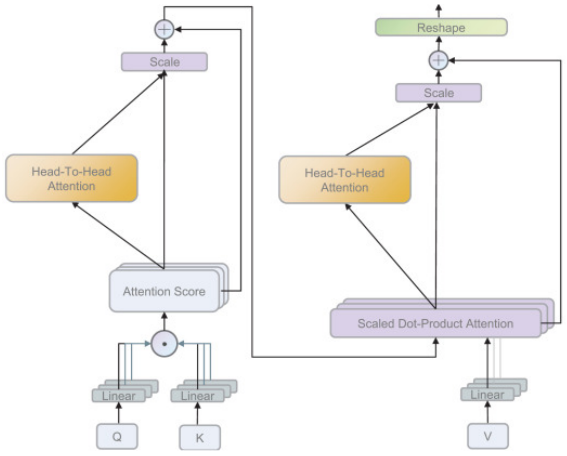

Figure 5: Estructura del mecanismo Head-to-Head Attention (HTHAttention) introducido en la arquitectura de RoBERTa. Fuente: Peng et al. (2024)

3.4. Modificación de Arquitectura Transformer

Varios estudios han recurrido a las modificaciones en la arquitectura Transformer para mejorar el rendimiento en tareas de clasificación de texto, aprovechando la potente capacidad del mecanismo de atención y las destacadas prestaciones de modelos como RoBERTa. Estos trabajos se centran en mejorar la capacidad de los modelos para manejar textos cortos, desbalanceo de clases y/o escasez de datos, manteniendo las características fundamentales de los Transformer, como la representación contextual proporcionada por el mecanismo de atención. Por ejemplo, Cheruku, Hussain, Kavati, Reddy, and Reddy (2023) proponen una modificación del modelo RoBERTa para extraer información contextualizada más relevante, combinándolo con redes neuronales recurrentes (RNN, Schmidt (2019)), con el objetivo de mejorar la clasificación de sentimientos en comentarios de Twitter. La propuesta incluye la adición de capas específicas para procesar metadatos y concatenar estos con las representaciones generadas por RoBERTa, tal como se ilustra en la Figura 4

Por otra parte, Tan, Lee, and Lim (2023) presentan una arquitectura híbrida entre RoBERTa y Gated Recurrent Units (GRU) Chung, Gulcehre, Cho, and Bengio (2014) para mejorar la clasificación de sentimientos, particularmente en datasets desbalanceados; de esta manera usando RoBERTa para generar *embeddings* discriminativos y GRU para capturar dependencias a largo plazo, permitió mejorar la capacidad de análisis de sentimientos, alcanzando una precisión de 91.28% en el conjunto de datos de Twitter utilizado. Finalmente, y siguiendo el enfoque de aumento de datos y el uso de frases auxiliares o prompts adoptado en este trabajo, Peng, Han, Zhong, Wu, and Zhang (2024) introducen una modificación en la arquitectura de RoBERTa, específicamente en su mecanismo de atención, a través de lo que denomina *Head-to-Head Attention* (HTHAttention). Este mecanismo, reflejado en la Figura 5, recalibra la importancia de cada cabeza de atención, mejorando la captura de relaciones complejas entre las distintas partes del texto y reduciendo redundancias. Además, el modelo incorpora *Prompt Text Augmentation* (PTA), una técnica que aumenta los datos mediante prompts o plantillas de texto con etiquetas embebidas, transformando la tarea de clasificación binaria (como positivo/negativo) en una tarea multitarea, permitiendo al modelo aprender a clasificar múltiples categorías simultáneamente y potenciando su capacidad de generalización.

3.5. Incorporación de Conocimiento mediante Frases Auxiliares

Otros de los principales desafíos es superar las limitaciones asociadas con la subjetividad de ciertas tareas y la necesidad de integrar información semántica más profunda para comprender el contexto. Estos problemas son especialmente relevantes en tareas como el análisis de sentimientos o la clasificación de texto en categorías complejas, como conspirativo o crítico, donde los matices lingüísticos y emocionales son relevantes para realizar clasificaciones precisas. En varios trabajos se ha propuesto un nuevo enfoque para mejorar el rendimiento de los modelos transformers mediante la introducción de información adicional, convirtiendo la tarea de clasificación tradicional en una clasificación en pares de oraciones (*sentence-pair classification*). Esta técnica consiste en tratar cada instancia como un par de oraciones que deben ser analizadas conjuntamente para predecir una etiqueta o categoría; en lugar de analizar una sola oración de forma aislada, el modelo recibe dos oraciones (que pueden estar relacionadas de diversas maneras, como una pregunta y su respuesta, o una oración y su contexto) y debe aprender la relación entre ellas para determinar su clasificación. Esta metodología ha demostrado ser una estrategia útil, particularmente cuando se complementa con el conocimiento del dominio. Por ejemplo, en Sun, Huang, and Qiu (2019) se consideran cuatro métodos diferentes para la construcción de estas frases auxiliares, entre los cuales se incluyen la formulación de preguntas (QA-M), la generación

Model	Aspect			Sentiment	
	Acc.	F_1	AUC	*Acc.*	AUC
LR	-	39.3	92.4	87.5	90.5
LSTM-Final	-	68.9	89.8	82.0	85.4
LSTM-Loc	-	69.3	89.7	81.9	83.9
LSTM+TA+SA	66.4	76.7	-	86.8	-
SenticLSTM	67.4	78.2	-	89.3	-
Dmu-Entnet	73.5	78.5	94.4	91.0	94.8
BERT-single	73.7	81.0	96.4	85.5	84.2
BERT-pair-QA-M	79.4	86.4	97.0	**93.6**	96.4
BERT-pair-NLI-M	78.3	87.0	**97.5**	92.1	96.5
BERT-pair-QA-B	79.2	**87.9**	97.1	93.3	**97.0**
BERT-pair-NLI-B	**79.8**	87.5	96.6	92.8	96.9

Table 5
Rendimiento con dataset SentiHood . Tabla comparativa del paper original Sun et al. (2019) con modelos de Saeidi, Bouchard, Liakata, and Riedel (2016), Ma, Peng, and Cambria (2018) y Liu, Cohn, and Baldwin (2018).

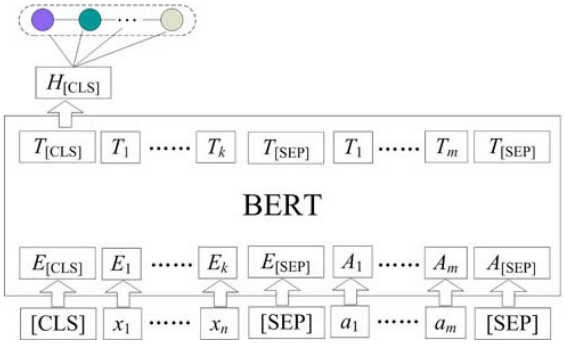

Figure 6: Uso de frases auxilires con modelo BERT. Fuente: Yu et al. (2019)

de pseudo-oraciones (NLI-M), y variaciones que incorporan la polaridad del sentimiento (QA-B y NLI-B). Estos métodos, ajustados al dominio, mejoraron significativamente el rendimiento del modelo, observando hasta una mejora de un 6% en las métricas de *accuracy* y de *F1-score* con respecto a un modelo BERT base (Véase la Tabla 5).

En Yu, Su, and Luo (2019) adoptaron esta estrategia presentando el modelo BERTT4TC que también aborda la escasez de datos y la disponibilidad de textos cortos, lo que le permitió lograr mejoras significativas en tareas multiclase y binarias superando a métodos tradicionales y otras variantes de BERT. Como se refleja en la Figura 6, se presenta la entrada original al modelo con la unión de este contexto adicional de diversas maneras diferenciadas entre sí por longitud y forma en la que se codifican las etiquetas (véase la Figura 7).

En tareas relacionadas con temas complejos, como puede ser la detección de discurso de odio o la clasificación de textos de peligro, Lin and Moh (2021) presentan modelos como COVID-Twitter-BERT (CT-BERT) que también emplea frases auxiliares para mejorar la clasificación de sentimientos en tweets relacionados con el COVID-19; aunque CT-BERT adaptado al dominio COVID-19, no

Model	Input Sequence	Label
BERT4TC-S	[CLS] I like this film. [SEP]	{negative, **positive**}
BERT4TC-AQ	[CLS] I like this film. [SEP] What is the result? [SEP]	{negative, **positive**}
BERT4TC-AA	[CLS] I like this film. [SEP] positive [SEP]	{0, **1**}
	[CLS] I like this film. [SEP] negative [SEP]	{**0**, 1}
BERT4TC-AWA	[CLS] I like this film. [SEP] The result is positive. [SEP]	{0, **1**}
	[CLS] I like this film. [SEP] The result is negative. [SEP]	{**0**, 1}

Figure 7: Diferentes métodos de uso de frases auxiliares con modelo BERT. Fuente: Yu et al. (2019)

siempre superó a BERT generalista, la integración de frases auxiliares demuestra un impacto positivo en el rendimiento del modelo, especialmente en escenarios complejos donde el conocimiento del dominio sobre la pandemia y su impacto emocional es clave. Un enfoque similar se emplea en la tarea de detección de spans tóxicos en Sánchez-Vega and Lopez-Monroy (2021), donde combinan información auxiliar directamente en las representaciones generadas por transformers mediante la concatenación de vectores de tokens con representaciones de oraciones completas (vector CLS) o combinaciones lineales de capas intermedias de BERT. Por último, usado un enfoque multimodal, Pijal, Armijos, Llumiquinga, Lalvay, Allauca, and Cuenca (2022) presentan el modelo CaTrBETO que utiliza frases auxiliares generadas a partir de imágenes asociadas a diferentes tweets, combinándolas con el texto original para clasificar sentimientos en español. Las frases auxiliares (Figura 8) se crean mediante otro modelo transformer que genera descripciones en inglés de las imágenes, las cuales luego son traducidas al español y enriquecidas con palabras clave relevantes del *tweet*.

Label	Tweet text	Auxiliar Sentence
0	Más crónicas rojas en Ecuador. En Salinas, Manta, por todos lados...	Un coche está aparcado al lado de la calle. inseguridad
0	Desde 2019 tenemos un debilitamiento del sistema penitenciario que a ...	Una mujer con traje y corbata CrisisCarcelaria
1	Tania Reneaum, secretaria Ejecutiva de la @CIDH, agradece por el informe...	Una mujer y un hombre sentados en una mesa con computadoras portátiles. CrisisCarcelaria
2	Recibimos en la Comisión de @SeguridadAN a los delegados de la..	Un grupo de personas sentadas en una mesa con computadoras portátiles. CrisisCarcelaria
1	#ECUADOR — Vigilia por #VictorGuaillas, dirigente de Molleturo que luchó...	Un grupo de personas de pie alrededor de una mesa con banderas. PenitenciariaDelLitoral

Figure 8: Ejemplos de frases auxiliares con modelo CaTrBETO. Fuente: Pijal et al. (2022)

Estas frases se integraron en un esquema de clasificación en pares de oraciones con BETO alcanzando una precisión del 60% y un F1-Macro de 0.70, superando las configuraciones tradicionales de un modelo BETO finamente ajustado y de un modelo BETO con TF-IDF (Tabla 6), demostrando la efectividad de la estrategia .

Model	Accuracy	Mac- F1
BETO FT	0,54	0,4280
BETO FT-IDF	0,56	0,2428
CaTrBETO	0,60	0,70

Table 6
Resultados del paper original del modelo CaTrBETO Pijal et al. (2022).

Class	CRITICAL	CONSPIRACY
Train	2,621	1379
Test	655	345

Table 7
Distribución de las clases del dataset

4. Metodología

En esta sección se explican los detalles del enfoque utilizado, véase la Figura 9. El diagrama refleja las etapas principales del proceso, que incluyen el aumento de datos, el preprocesamiento (*cleaning*, *normalization*, *tokenization*), el enriquecimiento de los datos mediante el análisis de emociones y la utilización de modelos de incrustación de palabras como BERT y RoBERTa, seguidos por la clasificación. Es importante señalar que no todas las etapas se aplican en todos los experimentos: en algunos casos no se realiza el aumento de datos, mientras que en otros se emplea validación *K-Fold* o *Stratified K-Fold*, o incluso se omiten estas técnicas, con el fin de comparar los resultados obtenidos.

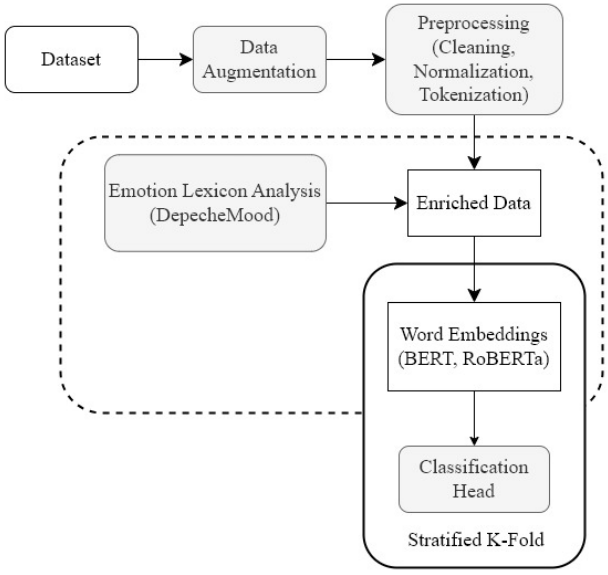

Figure 9: Diagrama de la metodología empleada

4.1. Dataset

Se emplea el dataset de pensamiento de oposición de la competición PAN at CLEF 2024, correspondiente a la tarea *Oppositional Thinking Analysis* de Korenčić et al. (2024). Esta tarea utiliza el corpus XAI-DisInfodemic, compuesto por 10,000 mensajes recopilados de 2,273 canales públicos de Telegram: 5,000 mensajes en inglés y 5,000 en español. Estos mensajes contienen puntos de vista oposicionales sobre la pandemia de COVID-19 y fueron obtenidos de canales públicos en los que los usuarios tienden a compartir mensajes contrarios al discurso dominante sobre la pandemia.

Los 5,000 mensajes por idioma se dividen en 4,000 mensajes para el conjunto de entrenamiento, publicado en la web oficial, y 1,000 mensajes para el conjunto de prueba, reservado para evaluar los modelos propuestos en la competición mediante un conjunto de datos no accesible ni manipulable por los participantes.

En cuanto a la longitud media de los textos, los mensajes en español tienen 128 tokens, mientras que los mensajes en inglés alcanzan un promedio de 265 tokens, siendo más extensos debido a una mayor tendencia a elaborar teorías conspirativas detalladas en este idioma.

Los textos están clasificados en dos categorías: mensajes críticos y mensajes conspiranoicos. Para crear este corpus, los autores desarrollaron un esquema de anotación que permite diferenciar entre mensajes que critican puntos de vista convencionales sobre el COVID-19 y aquellos que defienden la existencia de una conspiración.

Un mensaje es etiquetado como conspiranoico (CONSPIRATORY) si cumple al menos uno de los siguientes criterios:

- Identifica una estrategia de gestión de la pandemia como resultado de las acciones de una agencia secreta o un grupo reducido de personas con poder.
- Afirma que la pandemia no es real.
- Acusa a los críticos de las teorías conspirativas de ser parte de la trama.
- Divide a la sociedad en dos grupos: aquellos que "saben la verdad" (conspiranoicos) y los "ignorantes".

Por otro lado, un mensaje se clasifica como crítico (CRITICAL) si se opone a interpretaciones comúnmente aceptadas de los eventos relacionados con la pandemia, pero no muestra ninguna de las cuatro características asociadas a los mensajes conspiranoicos.

En la Tabla 7 se observa un desbalance entre las clases *CRITICAL* y *CONSPIRACY*, siendo la clase *CRITICAL* la mayoritaria con una diferencia de 1,242 ejemplos. Este desbalance puede afectar la precisión del modelo, ya que tiende a favorecer la clase mayoritaria, lo que podría reducir la efectividad en la clasificación de la clase minoritaria.

4.2. Modelos

En este trabajo, se emplearon modelos preentrenados de la familia BERT, específicamente **BERT base uncased**

Devlin et al. (2019) y **RoBERTa base** Liu et al. (2019), obtenidos del marco de trabajo Hugging Face. Ambos modelos fueron utilizados como base para establecer un modelo de referencia (*baseline*) y para evaluar si un enfoque alternativo basado en una modificación en la atención mejora el rendimiento.

Los modelos fueron *fine-tuned* utilizando los mismos parámetros de entrenamiento para asegurar una comparación justa:

- **Epochs:** 15
- **Learning rate:** 2e-5
- **Weight decay:** 0.01
- **Batch size:** 32

El tamaño máximo de secuencia para los tokens fue de 512, que es el límite estándar para estos modelos. Junto con los modelos, se utilizó el tokenizador correspondiente para transformar los datos crudos en secuencias de tokens, que es la unidad básica aceptada por los modelos de transformer. Este proceso de fine-tuning consistió en entrenar los modelos preentrenados con un conjunto de datos específico que no había sido visto previamente por los modelos.

4.3. Métricas

Para evaluar el rendimiento del modelo, se ha optado por utilizar la métrica **F1-Score macro** (3), dado su amplio uso en investigaciones similares y su capacidad para manejar situaciones de desequilibrio de clases. Esta métrica es particularmente útil ya que asigna un peso equilibrado a cada clase, independientemente de su frecuencia en el conjunto de datos, lo que permite una evaluación más justa del rendimiento del modelo en escenarios con clases desbalanceadas.

El F1-Score se calcula utilizando las métricas de precisión y recuperación, que se definen de la siguiente manera:

$$\text{Precisión} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Positivos}} \quad (1)$$

$$\text{Recuperación} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}} \quad (2)$$

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

El **F1-Score macro** se calcula como el promedio de los F1-Scores obtenidos para cada clase individualmente, sin tener en cuenta el desequilibrio entre las clases. Esto asegura que el modelo sea evaluado de manera justa, incluso si una clase es significativamente más prevalente que la otra. Este enfoque es esencial cuando se trabaja con datos desbalanceados, ya que un alto desempeño en la clase mayoritaria no debe ocultar un rendimiento deficiente en la clase minoritaria.

Además, para obtener una visión más completa del rendimiento del modelo, se utiliza el coeficiente **MCC** (Coeficiente de Correlación de Matthews), que es una medida más robusta en clasificación binaria, ya que toma en cuenta todas las posibles combinaciones de resultados (verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos). La fórmula para el MCC es la siguiente:

$$MCC = \frac{(VP \cdot VN) - (FP \cdot FN)}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \quad (4)$$

1. Numerador:
   - $(VP \cdot VN) - (FP \cdot FN)$: Representa el balance entre las predicciones correctas (positivos y negativos) y los errores cruzados. Este término mide cómo el modelo separa ambas clases.
2. Denominador: El producto de las combinaciones en el denominador captura toda la distribución de predicciones y valores reales. Esto garantiza que el MCC:
   - Penalice fuertemente los errores en cualquier clase, especialmente en la minoritaria (falsos negativos o falsos positivos).
   - Reconozca el desempeño en ambas clases de manera equitativa, incluso cuando los datos están desbalanceados.

El resultado de la métrica MCC se puede interpretar con tres casos distintos: +1 representa el mejor escenario, donde todas las predicciones son correctas. Un valor de 0 indica que el modelo no puede diferenciar entre las dos clases, funcionando de manera similar a un predictor aleatorio. Por último, un valor de -1 corresponde al peor caso, donde el modelo presenta un rendimiento totalmente inverso, es decir, siempre predice la clase incorrecta.

El MCC no se enfoca solo en una clase o subconjunto de métricas, sino que evalúa el equilibrio entre todas las combinaciones posibles de predicciones y valores reales, mitigando el desbalanceo en un conjunto de datos.

En este estudio, también se ha considerado el impacto del aumento de datos para equilibrar las clases, analizando cómo las variaciones en el F1-Score de cada clase reflejan los cambios en la distribución de los datos y el rendimiento general del modelo.

4.4. Experimentación

4.4.1. Preprocesado de Datos

Como parte del preprocesamiento de datos, se lleva a cabo una limpieza del texto eliminando enlaces, números y caracteres especiales, pero conservando los signos de puntuación, ya que pueden aportar información útil durante el entrenamiento del modelo. Además, el texto se normaliza convirtiéndolo a minúsculas, eliminando saltos de línea y

reduciendo múltiples espacios en blanco a uno solo, asegurando una estructura uniforme para su procesamiento posterior.

Una vez limpiados los datos, se utiliza el léxico de emociones para crear las frases auxiliares que servirán como contexto adicional para el modelo. Para asociar cada texto con una o más emociones, se emplea un método de coincidencia de palabras clave, presentes en el léxico DepecheMood++ de Araque, Gatti, Staiano, and Guerini (2018). Se emplean los lemas del léxico, contando la presencia de los lemas en cada texto y se calcula la media de los puntajes de emoción de los lemas coincidentes en el texto. En caso de que se asocie una sola emoción al texto, se selecciona la emoción con la probabilidad más alta. Un ejemplo del proceso se puede ver en la Figura 10.

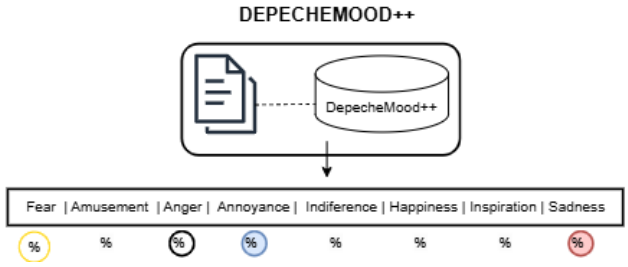

Figure 10: Uso del léxico DepecheMood++ para extraer las emociones reflejadas en el texto.

Para abordar el desbalance de clases, se implementan técnicas de aumento de datos descritas previamente. En particular, se utiliza un modelo preentrenado para traducir algunos textos del conjunto de entrenamiento al alemán y luego volver a traducirlos al inglés. Este proceso genera variaciones de los textos originales que son similares pero no idénticos, lo que contribuye a enriquecer el conjunto de datos. De este modo, se logra reducir la desigualdad entre clases, mejorando el equilibrio del conjunto de datos y potenciando los resultados del modelo en tareas de clasificación.

4.4.2. Fine Tuning

El proceso de *fine-tuning* se llevó a cabo utilizando modelos de transformers base BERT y RoBERTa, donde se ajustaron diferentes parámetros clave, como la tasa de aprendizaje y el tamaño del *batch*. Durante este ajuste, se exploraron distintas maneras de incorporar el contexto adicional para enriquecer la entrada de los modelos. Se probó la adición de emociones al inicio de la entrada, con el fin de aprovechar los mecanismos de atención de los transformers desde el principio.

En el caso de BERT, se utilizó la estructura original y se compararon dos enfoques para el manejo de la desigualdad de clases: *K-Fold* y *Stratified K-Fold*, con el objetivo de asegurar una mejor generalización en presencia de clases desbalanceadas. Por otro lado, para el modelo RoBERTa, y con los mejores parámetros empleados con BERT, se implementó una arquitectura modificada, añadiendo una cabeza de atención multi-cabeza (8 cabezas) para mejorar la capacidad del modelo de captar relaciones complejas entre las emociones y el contexto del texto, aumentando el rendimiento del modelo en la tarea de clasificación.

En ambos enfoques, se empleó un aumento de datos para ampliar la cantidad y diversidad del conjunto de entrenamiento, y se integraron las frases auxiliares generadas con las emociones asociadas a cada texto. Finalmente, se llevó a cabo un ajuste con RoBERTa-large para comparar el desempeño de los modelos base mejorados con el de un modelo de mayor tamaño y capacidad, lo que proporcionó una referencia de rendimiento para evaluar la efectividad de los enfoques implementados.

5. Resultados

Primero, se evaluó el desempeño del modelo utilizando los datos en su estado original, sin aplicar técnicas de ampliación. Esto permitió establecer una línea base inicial para medir el progreso y las mejoras posteriores. Posteriormente, se incorporó información adicional sobre emociones y se empleó la técnica de k-fold cross-validation para obtener una evaluación más robusta. Durante los experimentos, se probaron diferentes configuraciones de parámetros, logrando los mejores resultados con una longitud máxima de secuencia de 512 y una tasa de aprendizaje (learning rate) de 2e-5. Estos resultados se presentan en la Figura 11, donde también se incluyen los valores base (baselines) de cada métrica.

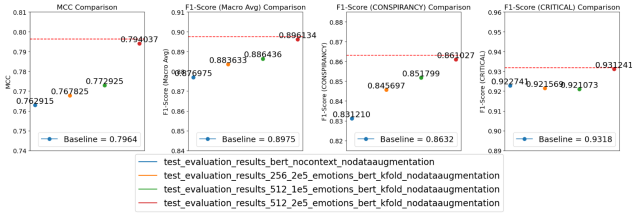

Figure 11: Desempeño del modelo BERT-base añadiendo emociones variando sus parámetros

Aunque no se alcanzaron los valores base, los resultados muestran una clara mejoría en comparación con el modelo inicial. Sin embargo, se observó una discrepancia notable en el F1-Score entre las diferentes clases, lo que motivó la aplicación de técnicas de aumento de datos. A pesar de esto, como se evidencia en la Figura 12, los resultados no fueron los esperados. Para abordar el desbalanceo de clases, se cambió a la técnica de *Stratified k-fold cross-validation*, que toma en cuenta este problema al dividir los datos. Aunque esta modificación mejoró algunas métricas en comparación con la evaluación previa, las mejoras no fueron significativas.

Finalmente, se optó por explorar un modelo más avanzado y con mayor capacidad: RoBERTa-large, con el objetivo de superar las limitaciones observadas con el modelo inicial. En una primera etapa, se utilizó este modelo incluyendo únicamente las emociones, siguiendo la misma

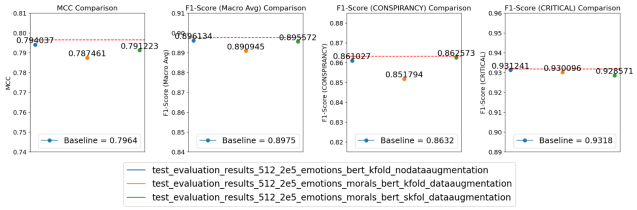

Figure 12: Desempeño del modelo BERT-base añadiendo emociones y aumentando datos

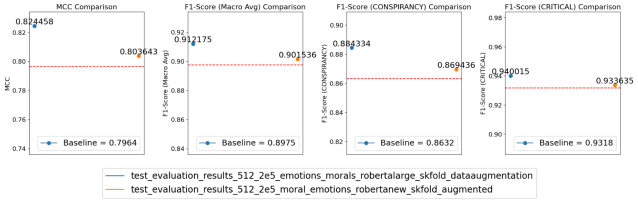

Figure 14: Desempeño del modelo RoBERTa modificado

metodología aplicada previamente con BERT. Los resultados se evaluaron comparando las técnicas de *K-fold* y *Stratified k-fold cross-validation* (*sk-fold*), mostrando un desempeño significativamente mejor con *sk-fold*. De hecho, este enfoque permitió superar los valores base en casi todas las métricas. Motivados por estos resultados, se decidió aplicar también técnicas de aumento de datos, manteniendo sk-fold como estrategia de evaluación. Esta combinación no solo reforzó el rendimiento del modelo, sino que permitió superar todos los valores base en las métricas evaluadas. La evolución y las mejoras alcanzadas se ilustran claramente en la Figura 13, destacando el impacto positivo del uso de RoBERTa-large junto con el aumento de datos y una estrategia de validación adecuada.

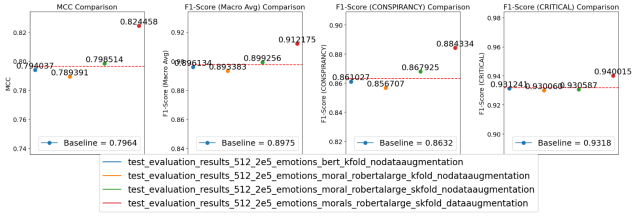

Figure 13: Desempeño del modelo RoBERTa-large

En nuestra exploración de técnicas avanzadas, realizamos modificaciones a la arquitectura de RoBERTa-base, incorporando capas de atención múltiple y convoluciones, siguiendo las directrices descritas en el artículo previamente citado. Estas adaptaciones permitieron superar el baseline, aunque los resultados obtenidos fueron superados por los alcanzados al aplicar nuestras estrategias sobre RoBERTa-large, probablemente debido a la superior capacidad y rendimiento inherentes a este modelo más avanzado. Apreciamos estos resultados en la Figura 14.

En la Tabla 15 se presenta una visión global de todos los resultados obtenidos para la métrica F1-Score. El valor de referencia *baseline* para esta métrica es 0.8975.

6. Conclusiones y Líneas Futuras

Los resultados obtenidos en este trabajo destacan la importancia de integrar información subjetiva, como el análisis de emociones, en modelos transformers para tareas que requieren captar matices complejos en el lenguaje,

		BERT-base	RoBERTa-large	RoBERTa modificado
Datos iniciales sin modificar		0.876975	-	-
Sólo con emociones	K-fold	0.896134	0.893383	-
	Sk-fold	-	0.899256	-
Con emociones y aumento	K-fold	0.890945	-	-
	Sk-fold	0.895572	0.912175	0.901536

Figure 15: Resumen de resultados obtenidos para la métrica F1-Score.

como la diferenciación entre textos conspirativos y críticos. Aunque los modelos base como BERT mostraron mejoras significativas con la incorporación de emociones y estrategias avanzadas de validación (como *Stratified k-fold cross-validation*), fue necesario recurrir a modelos más avanzados, como RoBERTa-large, para superar consistentemente los valores base en todas las métricas evaluadas. Esto pone de manifiesto las limitaciones de los modelos ligeros cuando se enfrentan a tareas de alta subjetividad con conjuntos de datos pequeños.

En nuestra exploración, comprobamos que técnicas como el aumento de datos y modificaciones arquitectónicas en modelos más ligeros pueden mitigar parcialmente estas limitaciones, pero su impacto no fue suficiente para alcanzar el rendimiento de modelos más grandes. RoBERTa-large, junto con un enfoque adecuado de validación y aumento de datos, permitió obtener resultados significativamente mejores, destacando su capacidad para aprovechar incluso pequeños volúmenes de datos cuando se emplean estrategias avanzadas.

Este trabajo demuestra que la incorporación de información subjetiva, junto con estrategias avanzadas de validación y aumento de datos, puede mejorar el desempeño de modelos transformers en tareas de alta subjetividad. Sin embargo, también subraya los desafíos de trabajar con conjuntos de datos pequeños y desbalanceados, así como la dependencia de modelos grandes para alcanzar un rendimiento óptimo. En el futuro, optimizar el uso de modelos ligeros y explorar formas innovadoras de integrar información subjetiva serán pasos clave para lograr un equilibrio entre precisión, eficiencia y escalabilidad.

6.1. Limitaciones

Una de las principales limitaciones fue la cantidad reducida de datos disponibles, aunque se aplicaron técnicas

de aumento de datos, estas no lograron compensar completamente la falta de diversidad en el conjunto de datos original. Por otra parte, modelos largos como RoBERTa-large plantean desafíos en términos aplicabilidad debido a los recursos limitados.

6.2. Trabajos Futuros

Como líneas futuras, se plantean enfoques diferentes, tales como:

1. Exploración de nuevas formas de integrar información subjetiva: más allá de incorporar emociones en las entradas mediante frases auxiliares, se propone explorar técnicas que integren esta información en las capas internas del modelo, como la fusión de los embeddings emocionales en capas de atención. Además, se plantea usar un enfoque multitarea, donde el modelo aprenda simultáneamente la tarea de clasificación principal y el análisis de emociones.
2. Evaluación de modelos alternativos: analizar el desempeño de otros modelos con arquitecturas distintas para contrastar los resultados obtenidos.
3. Ampliación del conjunto de datos: recopilar más datos relevantes y representativos para la tarea, con el objetivo de mejorar la generalización del modelo y su rendimiento en diferentes contextos.

7. Bibliografía

References

M. A. Hogg, Uncertain self in a changing world: A foundation for radicalisation, populism, and autocratic leadership, European Review of Social Psychology 32 (2021) 235–268.

D. Korenčić, et al., Overview of the oppositional thinking analysis pan task at clef 2024, in: Working Notes of CLEF, 2024.

H. Tajfel, An integrative theory of intergroup conflict, in: The social psychology of intergroup relations, Brooks/Cole, 1979, pp. 33–47.

K. M. Douglas, J. E. Uscinski, R. M. Sutton, A. Cichocka, T. Nefes, C. S. Ang, F. Deravi, Understanding conspiracy theories, Political psychology 40 (2019) 3–35.

M. L. Newman, J. W. Pennebaker, D. S. Berry, J. M. Richards, Lying words: Predicting deception from linguistic styles, Personality and social psychology bulletin 29 (2003) 665–675.

S. Kwon, M. Cha, K. Jung, W. Chen, Y. Wang, Prominent features of rumor propagation in online social media, in: 2013 IEEE 13th international conference on data mining, IEEE, 2013, pp. 1103–1108.

V. L. Rubin, N. Conroy, Y. Chen, S. Cornwell, Fake news or truth? using satirical cues to detect potentially misleading news, in: Proceedings of the second workshop on computational approaches to deception detection, 2016, pp. 7–17.

R. N. Zaeem, C. Li, K. S. Barber, On sentiment of online fake news, in: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2020, pp. 760–767.

N. Zhang, J. Song, K. Chen, S. Jia, Emotional contagion in the propagation of online rumors., Issues in Information Systems 23 (2022).

C. G. Horner, D. Galletta, J. Crawford, A. Shirsat, Emotions: The unexplored fuel of fake news on social media, in: Fake News on the Internet, Routledge, 2023, pp. 147–174.

T. Cosgrove, M. Bahr, The language of conspiracy theories: Negative emotions and themes facilitate diffusion online, Sage Open 14 (2024) 21582440241290413.

M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, J. Vilares, Sentiment analysis for fake news detection, Electronics 10 (2021) 1348.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

T. Pavlov, G. Mirceva, Covid-19 fake news detection by using bert and roberta models. in 2022 45th jubilee international convention on information, communication and electronic technology (mipro), 2022.

Y. Peskine, P. Papotti, R. Troncy, Detection of covid-19-related conpiracy theories in tweets using transformer-based models and node embedding techniques, in: MediaEval 2022, Multimedia Evaluation Workshop, 12-13 January 2023, Bergen, Norway, 2023.

A. Albladi, C. Seals, Detection of conspiracy vs. critical narratives and their elements using nlp, Working Notes of CLEF (2024).

Z. Liu, B. Liu, P. Thompson, K. Yang, S. Ananiadou, Conspemollm: Conspiracy theory detection using an emotion-based large language model, arXiv preprint arXiv:2403.06765 (2024).

A. Koufakou, J. Scott, Lexicon-enhancement of embedding-based approaches towards the detection of abusive language, in: Proceedings of the second workshop on trolling, aggression and cyberbullying, 2020, pp. 150–157.

J. Mutinda, W. Mwangi, G. Okeyo, Sentiment analysis of text reviews using lexicon-enhanced bert embedding (lebert) model with convolutional neural network, Applied Sciences 13 (2023) 1445.

S. K. P. Shrirang Mhalgi, S. Kübler, Iucl at pan 2024: Using data augmentation for conspiracy theory detection, in: IUCL at PAN 2024: Using Data Augmentation for Conspiracy Theory Detection, 2024. URL: https://downloads.webis.de/pan/publications/papers/mhalgi_2024.pdf.

M. C. A. Angelo Maximilian Tulbure, Conspiracy vs critical thinking using a ensemble of transformers with data augmentation techniques, in: Conspiracy vs Critical Thinking Using a Ensemble of Transformers with Data Augmentation Techniques, 2024. URL: https://downloads.webis.de/pan/publications/papers/tulbure_2024.pdf.

Helsinki-NLP, Helsinki-nlp/opus-mt-es-en, https://huggingface.co/Helsinki-NLP/opus-mt-es-en, 2024.

A. P.-L. Iñaki del Campo Sánchez-Hermosilla, D. Camacho, A study on nlp model ensembles and data augmentation techniques for separating critical thinking from conspiracy theories in english texts, in: A Study on NLP Model Ensembles and Data Augmentation Techniques for Separating Critical Thinking from Conspiracy Theories in English Texts, 2024. URL: https://downloads.webis.de/pan/publications/papers/sanchezhermosilla_2024.pdf.

nlpaug, nlpaug, 2024. URL: https://github.com/makcedward/nlpaug.

R. Cheruku, K. Hussain, I. Kavati, A. Reddy, K. Reddy, Sentiment classification with modified roberta and recurrent neural networks, Multimedia Tools and Applications 83 (2023) 1–19.

R. M. Schmidt, Recurrent neural networks (rnns): A gentle introduction and overview, 2019. URL: https://arxiv.org/abs/1912.05911. arXiv:1912.05911.

K. Tan, C.-P. Lee, K. Lim, Roberta-gru: A hybrid deep learning model for enhanced sentiment analysis, Applied Sciences 13 (2023) 3915.

J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. URL: https://arxiv.org/abs/1412.3555. arXiv:1412.3555.

B. Peng, K. Han, L. Zhong, S. Wu, T. Zhang, A head-to-head attention with prompt text augmentation for text classification, Neurocomputing 595 (2024) 127815.

C. Sun, L. Huang, X. Qiu, Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 380–385. URL: https://aclanthology.org/N19-1035. doi:10.

18653/v1/N19-1035.

M. Saeidi, G. Bouchard, M. Liakata, S. Riedel, SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods, in: Y. Matsumoto, R. Prasad (Eds.), Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 1546–1556. URL: https://aclanthology.org/C16-1146.

Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm, 2018.

F. Liu, T. Cohn, T. Baldwin, Recurrent entity networks with delayed memory update for targeted aspect-based sentiment analysis, 2018. URL: https://arxiv.org/abs/1804.11019. arXiv:1804.11019.

S. Yu, J. Su, D. Luo, Improving bert-based text classification with auxiliary sentence and domain knowledge, IEEE Access 7 (2019) 176600–176612.

H. Y. Lin, T.-S. Moh, Sentiment analysis on covid tweets using covid-twitter-bert with auxiliary sentence approach, in: Proceedings of the 2021 ACM Southeast Conference, ACMSE '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 234–238. URL: https://doi.org/10.1145/3409334.3452074. doi:10.1145/3409334.3452074.

F. Sánchez-Vega, A. P. Lopez-Monroy, Bert's auxiliary sentence focused on word's information for offensiveness detection, in: IberLEF@SEPLN, 2021. URL: https://api.semanticscholar.org/CorpusID:238207249.

W. Pijal, A. Armijos, J. Llumiquinga, S. Lalvay, S. Allauca, E. Cuenca, Spanish pre-trained catrbeto model for sentiment classification in twitter, in: 2022 Third International Conference on Information Systems and Software Technologies (ICI2ST), 2022, pp. 93–98. doi:10.1109/ICI2ST57350.2022.00021.

O. Araque, L. Gatti, J. Staiano, M. Guerini, Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques, 2018. URL: https://arxiv.org/abs/1810.03660. arXiv:1810.03660.