ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.1076

News Categorisation Based on Pre-Trained Transformer Models

César Espin-Riofrio, MSc.[1], Vanessa Murillo-Cepeda, Ing.[1], David García-Zambrano, Ing.[1], Verónica Mendoza Morán, MSc.[1], Johanna Zumba Gamboa, MSc.[1], Arturo Montejo-Ráez, PhD.[2]
[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, vanessa.murilloc@ug.edu.ec, davidf.garciaz@ug.edu.ec, *veronica.mendozam@ug.edu.ec*, johanna.zumbag@ug.edu.ec
[2]Universidad de Jaén, España, amontejo@ujaen.es

Abstract- The rise of digital journalism, the amount of news and the continuous number of people accessing these contents, often generates that third parties, through web platforms and social networks, have the opportunity to persuade readers with content that alters their opinion or behaviour on a topic, so it is necessary to classify news using Natural Language Processing (NLP) techniques. This work seeks to experiment with pre-trained Transformer models using transfer learning and fine tuning to obtain a model capable of determining whether a news item is satire, opinion or information. To do so, we use a labelled dataset of news in English presented for the SemEval 2023 campaign, translating it into Spanish to experiment also in this language. We use pre-trained Transformer models for text classification tasks in the mentioned languages, thus, we compare several models and their predictions using evaluation metrics. The results give indications of the goodness of the models considering the subjective type of news, in the case of satire and opinion, and objective for information, thus contributing to future research related to text classification, specifically news categorisation.

Keywords- Natural Language Processing, Transformer models, news categorisation.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

Categorización de Noticias Basado en Modelos Pre-Entrenados Transformer

César Espin-Riofrio, MSc.[1], Vanessa Murillo-Cepeda, Ing.[1], David García-Zambrano, Ing.[1], Verónica Mendoza Morán, MSc.[1], Johanna Zumba Gamboa, MSc.[1], y Arturo Montejo-Ráez, PhD.[2]
[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, vanessa.murilloc@ug.edu.ec, davidf.garciaz@ug.edu.ec, *veronica.mendozam@ug.edu.ec*, johanna.zumbag@ug.edu.ec
[2]Universidad de Jaén, España, amontejo@ujaen.es

Resumen– ***El auge del periodismo digital, la cantidad de noticias y el continuo número de personas que acceden a estos contenidos, genera muchas veces que terceros por medio de las plataformas webs y redes sociales tengan la oportunidad de persuadir a los lectores con contenido que altere su opinión o comportamiento sobre un tema, por esto resulta necesario la clasificación de noticias mediante técnicas de Procesamiento de Lenguaje Natural (PLN). Este trabajo busca experimentar con modelos pre entrenados Transformer haciendo uso de aprendizaje por transferencia y ajuste fino para obtener un modelo capaz de determinar si una noticia es de tipo sátira, opinión o información. Para ello usamos un dataset etiquetado de noticias en inglés presentado para la campaña SemEval 2023, traduciéndolo al español para experimentar también en este idioma. Utilizamos modelos Transformer pre entrenados para tareas de clasificación de textos en los idiomas mencionados, así, comparamos varios modelos y sus predicciones mediante métricas de evaluación. Los resultados dan indicios de las bondades de los modelos considerando el tipo de noticia subjetiva, en el caso de sátira y opinión, y objetiva para información, contribuyendo así a futuras investigaciones relacionadas a la clasificación de textos en específico la categorización de noticias.***

Palabras claves— ***Procesamiento de Lenguaje Natural, modelos Transformer, categorización de noticias.***

I. Introducción

En la actualidad el internet es casi esencial en la vida de las personas debido a su aporte en las actividades rutinarias como el trabajo, estudio, ocio, etc., que cada vez aumenta el número de conectados a la red. Esto ha permitido que la información que se comparte por este medio aumente a pasos agigantados y lleve a diversos negocios, como el periodismo multimedia, que aprovechen de las nuevas tendencias de consumo y se adapten para continuar llegando a las personas, dando origen al periodismo digital, donde se comparten noticias a diario y a su vez genera una gran cantidad de datos constantemente. [1] indica que, en los últimos años el enorme incremento en el uso del internet ha permitido el auge del periodismo multimedia y que la tendencia de uso de dispositivos inteligentes como teléfonos móviles, laptops, computadoras y tablets es cada vez mayor como alternativa a los periódicos. Una gran ventaja de esto es que se puede acceder a la información disponible en cualquier momento y horario, a diferencia del periódico tradicional, que había un tiempo de espera para poder adquirir una nueva edición impresa.

Con frecuencia se encuentran noticias en diferentes sitios como redes sociales o páginas web sin saber su tipo o no se logra identificar si se trata de un artículo de opinión, de información o de sátira. Para [2] la sátira es expresada en lenguaje figurativo, que por medio de la ironía y el humor se ridiculiza o se critica un evento o entidad, por lo que pueden llegar a ser engañosas y dañinas. Junto con los artículos de opinión, estos son subjetivos ya que son generados desde una perspectiva o punto de vista

En la actualidad las noticias tienen tal importancia en la opinión de las personas que pueden generar un impacto positivo o negativo, ya sea hacia un producto, partido político o compañía [3]. Según [4] las noticias falsas pueden promover ideas que favorecen a ciertas personas y desacreditan o desfavorecen a otras, ya que estas pueden estar sesgadas para manipular, en especial los artículos de tipo satírico.

Categorizar una noticia de forma tradicional lleva emplear una gran cantidad de tiempo, ya que la persona encargada tendría que hacer una lectura previa para luego tomar una decisión, esto sumado al incremento diario de noticias lo que vuelve más complejo el proceso.

Como consecuencia, las noticias originan muchas veces desinformación llegando a influir en la conducta o comportamiento de las personas, ya que pueden ser sesgadas y no siempre lo que se publica es de manera objetiva. Además, la falta de herramientas o técnicas que permitan la categorización de noticias impide que las plataformas digitales se encuentren ordenadas para facilitar a los usuarios información de fuentes objetivas.

Entonces, es necesario categorizarlas de forma automática sin que tome tanto tiempo como de forma manual, y que estén clasificados de tal manera que se evite la desinformación y la persuasión, es aquí donde entra el uso de técnicas de Procesamiento de Lenguaje Natural (PLN), [5] indican que se utilizan para resolver tareas automáticamente que necesitarían de personas para ser llevadas a cabo, tales como reconocimiento de entidades, generación de texto, traducción automática, etc.

La finalidad de esta investigación es categorizar noticias en inglés y español utilizando mecanismos de autoatención Transformer para determinar si es un artículo de opinión, información o sátira, partiendo de establecer la base conceptual para la categorización de noticias, experimentar realizando ajuste fino (fine-tuning) de modelos pre entrenados para

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

obtener nuevos modelos y finalmente presentar los resultados comparativos de predicción mediante métricas de evaluación, logrando determinar la categoría a la que pertenece una noticia.

Entre las distintas aportaciones hechas en lo que respecta a categorización de noticias mediante técnicas de clasificación encontramos a [6] que experimentan con el modelo de red neuronal convolucional bidirectional Long Short-Term Memory (LSTM) con el fin de mejorar la precisión de clasificación con respecto a otros métodos como Frecuencia de Término Frecuencia Inversa de Documento (TF-IDF), Support Vector Machine (SVM), Convolutional Neural Network (CNN) y LSTM. [7] presentaron la aplicación del modelo Bernoulli para categorizar noticias, logrando una tasa de exactitud del 98.4%. Existen estudios aplicados a otros idiomas, [8] evaluaron la clasificación para el nepalí con diversos algoritmos de ML, en el que el más apto fue Radial Basis Function (RBF) Kernel SVM. [9] Realizaron un sistema de clasificación de textos de aprendizaje automático usando Regresión Logística (LR), Random Forest (RF) y K-nearest Neighbor (KNN) donde LR logró la tasa de precisión más elevada. Por otro lado, [10] propusieron el uso de modelos Transformer para la clasificación de noticias aplicado al idioma portugués, el modelo BERTimbau superó a los demás modelos con los que se lo comparó.

Para [11] los modelos Transformer identifican las reglas gramaticales y la semántica para una mejor comprensión del texto, frase o palabra dentro de una oración. BERT [12] es muy utilizado para clasificación de textos actualmente. [13] Analizaron SVM, Naive Bayes y Random Forest y BERT, donde este último obtuvo la mejor precisión. En cuanto a la clasificación de noticias para diferenciar entre la categoría falsa y sátira, [14] usaron el modelo Transformer DistilBERT, mediante un ajuste fino de DistilBERT-base-uncased, con el que obtuvieron una mejor precisión con respecto a otros modelos de aprendizaje profundo. [15] Realizaron un análisis cuantitativo entre las variantes de SVM, que son: Twin Support Vector Machine (TWSVM), Least Square Support Vector Machines (LS-SVM) y Least Squares Twin Support Vector Machine (LS-TWSVM) en el que esta última logra mejor resultados en precisión y en tiempo de entrenamiento y prueba.

Con respecto a la clasificación de textos en idioma español, [16] utilizaron el modelo Transformer Selectra-Medium en el que lograron la clasificación y etiquetado de textos cortos en español provenientes de tweets.

[17] realizan la predicción de la complejidad de palabras simples en el idioma español obteniendo resultados tras ejecutar los modelos afinados basados en Transformer y ejecutados sobre los modelos pre entrenados BERT, XLM-RoBERTa, y RoBERTa-large-BNE y ejecutados sobre varios algoritmos de regresión.

Hay varios conceptos claves que sustentan nuestra investigación tales como el de ML según [18] son algoritmos que mediante el procesamiento de datos que se le proporciona, son capaces de extraer información útil para realizar tareas de forma automática que requieren de conocimiento humano. Una técnica en la que un modelo aprende es el aprendizaje supervisado que consiste en brindar datos con los resultados esperados, denominados etiquetas, de esta manera en el entrenamiento la máquina basará su aprendizaje con el etiquetado que se le proporciona [19]. Es muy utilizado en tareas de predicciones futuras que se basan en comportamientos o características vistas anteriormente para así buscar patrones relacionando todos los campos con uno específico, el cual se llama campo objetivo, y así obtener la salida correcta [20].

La clasificación de texto se basa en utilizar datos textuales ya sea una palabra, frase, párrafo o un documento, estos textos son sucesos que están interpretados en forma de textos, que sirven de entrenamiento para identificar características o patrones útiles para hacer clasificaciones [21]. Proveniente de esto tenemos la categorización de noticias, que consiste en la agrupación de noticias de tal forma que puedan simplificar la navegación del usuario, estas categorías basan su clasificación dependiendo de su contenido ya sea del título como del cuerpo, se considera una actividad compleja ya que requiere de varios pasos [22]. En la clasificación la variable por predecir es un conjunto de estados discretos o categórico, que a su vez pueden ser binarios, múltiples u ordenadas, donde trata de encontrar patrones en los datos y los clasifica en grupos, luego compara los nuevos datos y los ubica en uno de los grupos y así determinar de qué se trata [23].

Hugging Face[1] aloja arquitecturas basadas en Transformer, ofrece modelos previamente entrenados listos para ser probados y experimentados o también para compartirlos con finalidades investigativas o para producción. Los modelos Transformer [24] surgieron para solventar la dificultad que poseían los modelos recurrentes, que al tener que calcular las posiciones lo hacían de manera secuencial de forma que puedan mantener el contexto de la palabra anterior de un texto. El problema se manifestaba cuando las secuencias eran más grandes que la memoria, y esto puede llegar a ser una limitante. Los Transformer poseen la característica de paralelización, que se centra en los mecanismos de atención siguiendo una estructura de codificador y decodificador que están conectados y procesan los elementos en paralelo gracias a que en sus capas se dividen por dos subcapas: una es la de autoatención y la otra de retroalimentación que les permiten llevar la posición de cada elemento.

Su evolución comienza con Attention Is All You Need [25] dando solución a problemas como el manejo de textos largos y superando a modelos como RNN y LSTM. BERT cuenta con una arquitectura que le permite entrenar de manera bidireccional con texto sin etiquetar, por lo que es capaz de analizar las palabras que conforman el texto de izquierda a derecha tomando como referencia una palabra clave y de esta manera puede extraer el contexto. [26] presentaron ULMFit, una técnica de fine-tuning basado en Universal Language

[1] https://huggingface.co/

Model (ULM) para mejorar el rendimiento en clasificación de texto mediante adaptación fina, mediante el uso de las técnicas discriminative fine-tuning y slanted triangular learning rates.

Generative Pre-Training (GPT) demostró que se pueden obtener buenos resultados en el campo del PLN con entrenamiento de textos sin etiquetar, aplicando luego un ajuste fino discriminativo para cada tarea especifica, permitiendo realizar con éxito diversas tareas como responder preguntas, evaluación de igual semántica y clasificación de texto [27]. Su segunda versión, GPT-2 puede desenvolverse en diversas tareas, tales como, repuesta de preguntas y generación de resúmenes. Utiliza la gran cantidad de datos que aprendió previamente para generar de manera automática entradas y etiquetas en base a estos textos [28]. GPT-3 supera a los modelos antecesores en 10 veces más parámetros, las pruebas de rendimiento que hicieron fueron con pocos disparos, que les dio resultados positivos en tareas de traducción, respuestas a preguntas, además de tareas que necesitan razonamiento sobre la marcha como, descifrar palabras o usar una palabra nueva dentro de una oración y generar muestras de noticias similar a como lo harían las personas [29].

Transformer-XL es un modelo capaz de manejar contextos más largos en comparación con la arquitectura inicial de Transformer, esto es posible gracias a que cuenta con un mecanismo de recurrencia a nivel de segmento y un nuevo esquema de codificación posicional, este modelo logra generar artículos de texto coherentes que contienen miles de tokens [30].

XLNet, que significa entrenamiento previo autorregresivo generalizado para la comprensión del lenguaje, es un modelo enfocado en mejorar los problemas de los modelos basados en la atención, como BERT, utilizando la técnica de autoagresión condicional logrando posibles cambios de las palabras en una oración para una mejor comprensión del texto [31]. Con la ayuda de Pytorch-Transformer, una librería basada en Pytorch permite utilizar los modelos de procesamiento basados en Transformer, tales como BERT, XLNet, GPT, entre otros. Se puede personalizar los modelos de una forma fácil para adaptarlos a sus tareas específicas [32].

Spacy-Pytorch-Transformer es una biblioteca que proporciona pipelines de los modelos spaCy que agrupa el paquete de pytorch-transformer proveniente de Hugging Face para que se puedan usar en spaCy [33]. XLM [34] es un modelo que utiliza la atención para detectar patrones en los datos y las relaciones entre ellos. Tiene la capacidad de trabajar con cualquier idioma y se puede adaptar a tareas específicas de fine-tuning, logrando una mejor comprensión del texto en comparación a otros modelos pre entrenados como BERT.

BART [35] utiliza técnicas de ruido arbitrario para corromper el texto y a partir de ahí construye el texto original, con este modelo obtuvieron resultados positivos en tareas de PLN por su compresión de texto que le permite generar texto, respuestas a preguntas y traducción de idiomas. Finalmente, Text-to-Text Transfer Transformer (T5) es un modelo codificador-decodificador entrenado en una mezcla multitarea de tareas supervisadas y no supervisadas, donde cada tarea se convierte en un formato de texto a texto como mejora para realizar tareas de PLN sin tener la necesidad de hacer un fine-tuning para cada tarea específica [36].

En esta investigación experimentamos con varios modelos Transformer pre entrenados para clasificación de textos en inglés y en español, según se aprecia en la Tabla 1.

TABLA 1
MODELOS UTILIZADOS

Modelo	Idioma
DistilBERT-base-uncased	Inglés
RoBERTa-base	Inglés
BERT-base-spanish-wwm-cased	Español
RoBERTa-base-BNE	Español
IXAmBERT-base-cased	Español
XLM-RoBERTa-base	Español
mDeBERTa-v3-base	Español

DistilBERT-base-uncased [37] es más rápido y pequeño que BERT, fue destilado a partir del checkpoint (punto de control) de bert-base-uncased, tiene 6 capas, 768 de tamaño oculto, 12 cabezas de atención y 66 millones de parámetros, "uncased" hace referencia a que fue entrenado con textos en minúsculas [38]. RoBERTa-base [39] es una configuración base de RoBERTa, los hiperparámetros de preentrenamiento que usa son 12 número de capas, 12 cabezas de atención y 8K de tamaño de lote (batch size), a comparación de BERT el rendimiento de este modelo mejora en tareas posteriores cuando aumenta la cantidad y la diversidad de los datos de entrenamiento.

Por parte de los modelos multilingües está BERT-base-spanish-wwm-cased [40] también conocido como Spanish-BERT, es un modelo de lenguaje basado en BERT que está entrenado específicamente para el idioma español, con 110 millones de parámetros mediante el uso de la técnica de enmascaramiento de palabras completas. El corpus usado es de un aproximado de 3 mil millones de palabras de fuentes como Wikipedia e información contenida en revistas gubernamentales, noticas, etc., este modelo es capaz de lograr mejores resultados en tareas en español que los otros modelos de multilenguaje que parten de BERT. Asimismo, IXAmBERT-base-cased [41] es un modelo basado en BERT, inicialmente surgió para la tarea de respuesta a preguntas para el idioma euskera, el corpus con el que fue entrenado está formado por bibliografías de Wikipedia en euskera más la Wikipedia en inglés y en español con 2,5M y 650M tokens respectivamente, por lo que puede ser aplicado para diversas tareas de procesamiento de lenguaje natural en español. En cambio, RoBERTa-base-BNE forma parte de los modelos MarIA [42], que están especializados para el idioma español. El corpus con el que han sido entrenados es de 507GB de textos limpios y deduplicados, que contienen alrededor de 135 mil millones de palabras las cuales fueron extraídas del Archivo Web del Español construido por la Biblioteca Nacional de

España (BNE) entre los años 2009 y 2019. Este modelo está basado en la arquitectura de RoBERTa-base que consta con 12 capas, 768 ocultas, 12 cabezas de atención y 125M de parámetros. De igual manera, XLM-RoBERTa-base [43] está basado en la versión base de RoBERTa denominada también XLM-R Base, fue entrenado con 2.5 TB de manera auto supervisada, la configuración usada para el entrenamiento consistió en 12 capas, 768 ocultas, 12 atenciones multi cabeza y 270M de parámetros, donde se usó un tamaño de vocabulario de 250K [43]. Finalmente, mDeBERTa-v3-base [44] que es una mejora de la versión original DeBERTa [45] que está a su vez mejora con respecto al modelo BERT y RoBERTa porque utiliza atención desenredada y decodificador de máscara mejorado, mDeBERTa es un modelo multilingüe con un conjunto de datos de 2.5T, cuenta con un vocabulario de 250k tokens su configuración es de 12 capas, 768 tamaños ocultos y 12 cabezas de atención, su rendimiento fue evaluado en 15 idiomas.

La presente investigación presenta en su contenido la metodología que se aplicó, el dataset utilizado junto a las técnicas para llevar a cabo el procesamiento y balanceo de datos, se describe el proceso de experimentación para obtener modelos de clasificación a partir de ajuste fino, se hizo un análisis comparativo con los modelos obtenidos, a través, de métricas de evaluación, para determinar el desempeño de cada uno.

II. Metodología

La metodología de esta investigación es bibliográfica documental, mediante la cual se realizó un recorrido de los modelos Transformer a través de la investigación de artículos científicos de impacto enfocados al tema propuesto. En ese mismo sentido, se determinaron varios modelos de categorización para realizar pruebas con ellos, por lo que, se determina que esta investigación también es cuasi experimental, ya que se utilizaron los distintos modelos seleccionados con el fin de analizar el rendimiento de estos y compararlos entre ellos. A su vez, es de tipo cuantitativa, puesto que los resultados serán evaluados de acuerdo con las métricas de evaluación.

A continuación, en la Fig. 1 se muestra el esquema de experimentación de los modelos.

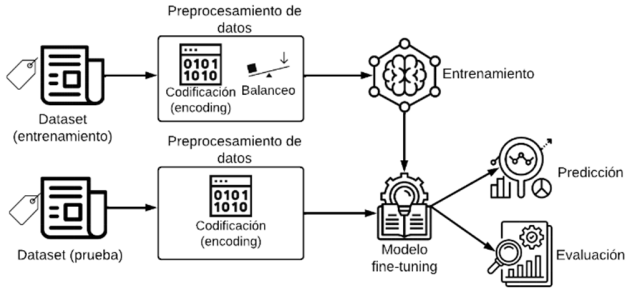

Fig. 1 Esquema del proceso de experimentación.

[2] https://propaganda.math.unipd.it/semeval2023task3/

A. *Dataset utilizado*

El dataset que se utilizó para el entrenamiento y prueba de los modelos fue tomado de la tarea tres de SemEval 2023[2]. Los conjuntos de datos contienen artículos en seis idiomas, de los cuales se ha tomado en cuenta los de idioma inglés que están compuestos por información recopilada desde 2020 hasta mediados de 2022 de fuentes como Google News y Europe Media Monitor (EMM). Está dividido en entrenamiento y prueba. El de entrenamiento cuenta con 433 muestras respectivamente etiquetadas en sátira, opinión e información, mientras que el dataset de prueba cuenta con 83 muestras, como se presenta en la Tabla 2.

TABLA 2
CANTIDAD DE MUESTRAS DE LOS DATASETS

Dataset	Cantidad de muestras	
Entrenamiento	Sátira	10
	Opinión	382
	Información	41
	Total	**433**
Prueba	Sátira	7
	Opinión	64
	Información	12
	Total	**83**

Finalmente se hizo una traducción al español de ambos conjuntos de datos para poder realizar la experimentación de los modelos también en este idioma, tal como se puede apreciar en la Fig. 2.

	id	text	label
0	833042063	Chelsea Handler admite que se siente "muy atra...	satire
1	832959523	Cómo Theresa May hizo una chapuza\n\nQué tiemp...	satire
2	833039623	Robert Mueller III descansa su caso - los demó...	satire
3	833032367	Robert Mueller no recomienda más acusaciones\n...	satire
4	814777937	La extrema derecha intenta cooptar a los chale...	satire
5	821744708	"Un lugar especial en el infierno" para quiene...	satire
6	833036489	Bill Maher dice que no necesita el informe Mue...	satire
7	707566605	Brote en Madagascar: Es "inevitable" que la pe...	opinion
8	708561738	¿Qué le parece que el Congreso pague las indem...	satire
9	782086447	El ex nuncio apostólico en Estados Unidos acus...	opinion
10	710376094	Un asteroide pasará rozando la Tierra días ant...	opinion
11	754179642	Un nuevo brote de ébola causa 17 muertos en el...	opinion
12	757243988	Una escuela secundaria de Virginia Occidental ...	reporting
13	729303442	La votación contra un político estudiantil jud...	opinion
14	699478811	Se intensifica la agresión iraní\n\nEl pasado ...	opinion

Fig. 2 Dataset de entrenamiento en español sin preprocesar.

B. *Preprocesamiento de datos*

En esta etapa se realizó una codificación categórica en los datasets de entrenamiento y de prueba de las etiquetas en la que cada una es representada por un número: 0 para sátiras, 1 para opinión y 2 para información, ya que para que un modelo pueda procesar las categorías, es necesario convertir las entradas en números o binarios. Se puede apreciar el resultado en la Fig. 3.

	text	label
0	Chelsea Handler admite que se siente "muy atra...	0
1	Cómo Theresa May hizo una chapuza\n\nQué tiemp...	0
2	Robert Mueller III descansa su caso - los demó...	0
3	Robert Mueller no recomienda más acusaciones\n...	0
4	La extrema derecha intenta cooptar a los chale...	0
5	"Un lugar especial en el infierno" para quiene...	0
6	Bill Maher dice que no necesita el informe Mue...	0
7	Brote en Madagascar: Es "inevitable" que la pe...	1
8	¿Qué le parece que el Congreso pague las indem...	0
9	El ex nuncio apostólico en Estados Unidos acus...	1
10	Un asteroide pasará rozando la Tierra días ant...	1
11	Un nuevo brote de ébola causa 17 muertos en el...	1
12	Una escuela secundaria de Virginia Occidental ...	2
13	La votación contra un político estudiantil jud...	1
14	Se intensifica la agresión iraní\n\nEl pasado ...	1

Fig. 3 Dataset en español con etiquetas codificadas.

Se debe tomar en cuenta que el dataset que se utiliza en el entrenamiento presenta un desbalanceo de clases, para tratarlo se usó el método de cálculo de pérdida CrossEntropyLoss[3] el cual recibe como parámetro los pesos asignados (weight) para cada clase, asignado mayor peso según menos cantidad de datos tenga, la fórmula (1) se usó para dicho cálculo.

$$\frac{(n - n_i)}{(m - 1) * n} \qquad (1)$$

n: Total de ejemplos en el corpus.
n_i*:* Número de ejemplos de la clase i.
m: Número de clases diferentes.

C. *Experimentación*

Una vez codificados los datasets de entrenamiento y prueba, se procedió con la experimentación de los modelos pre entrenados, los cuales mediante el aprendizaje por transferencia se pueden adaptar a tareas específicas, a través de un ajuste fino o fine-tuning. Para las pruebas se utilizó el dataset en inglés con los modelos de DistilBERT-base-uncased y RoBERTa-base, y para el idioma español BERT-base-spanish-wwm-cased, RoBERTa-base-BNE, IXAmBERT-base-cased, XLM-RoBERTa-base y mDeBERTa-v3-base. Para determinar sus rendimientos en la categorización de noticias, se lo hizo por medio de técnicas de evaluación, como las métricas de matriz de confusión, precisión, recall y F1-score.

D. *Tokenización*

Como los modelos necesitan de valores numéricos para procesar la información, se utilizó el tokenizador respectivo recomendado para cada modelo, como se observa en la Tabla 3.

TABLA 3
TOKENIZADORES USADOS

Modelo	Tokenizador
DistilBERT-base-uncased	DistilBertTokenizer
RoBERTa-base	RobertaTokenizer
BERT-base-spanish-wwm-cased	AutoTokenizer
RoBERTa-base-BNE	RobertaTokenizer
IXAmBERT-base-cased	AutoTokenizer
XLM-RoBERTa-base	AutoTokenizer
mDeBERTa-v3-base	AutoTokenizer

Se procede a hacer uso de estos con el fin de convertir los textos provenientes del dataset de noticia en representaciones numéricas para que puedan ser procesados por los modelos para su entrenamiento. Debido a las limitaciones de longitud soportada por los modelos escogidos, se procedió a dividir por secuencias de textos cada noticia para que puedan ser tokenizadas y que estas no superen la longitud que permiten los modelos. Como se trató de no superar los 512 tokens, se modificó la secuencia en las que venían dividas las noticias, por lo que se obtuvieron distintas cantidades de particiones en cada caso.

Una vez determinadas las divisiones que se obtuvieron de las noticias se procede a aplicar el cálculo para los pesos respectivos a cada caso con (1), estos se detallan en la Tabla 4.

TABLA 4
PESOS POR CLASE DE LOS DATASETS

Modelo	Pesos obtenidos		
	Sátira	Opinión	Información
DistilBERT-base-uncased	0.4919	0.0572	0.4508
RoBERTa-base	0.4919	0.0572	0.4508
BERT-base-spanish-wwm-cased	0.4916	0.0580	0.4504
RoBERTa-base-BNE	0.4916	0.0584	0.4501
IXAmBERT-base-cased	0.4921	0.0580	0.4500
XLM-RoBERTa-base	0.4920	0.0574	0.4506
mDeBERTa-v3-base	0.4919	0.0587	0.4494

E. *Fine-Tuning*

Obtenidos los datos tokenizados, se procedió a cargar el respectivo modelo utilizando CUDA para potenciar las capacidades de procesamiento de la GPU del entorno de trabajo Google Colab, luego se procedió con el fine-tuning de modo que se pueda obtener modelos de predicción de categorización de noticias. En Fig. 4 se observa la carga del modelo pre entrenado distilbert-base-uncased.

```
from_pretrained("distilbert-base-uncased", num_labels=3)
```

Fig. 4 Carga del modelo distilbert-base-uncased.

F. *Definición de hiperparámetros*

Para todos los modelos se aplicó los mismos hiperparámetros con la finalidad de realizar una comparación

[3] https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

uniforme, estos son tamaño de lote (batch size), epochs y tasa de aprendizaje, indicados en Tabla 5.

TABLA 5
HIPERPARÁMETROS USADOS EN LOS MODELOS

Hiper parámetro	Valor
Tamaño de lote (batch size)	8
Epochs	5
Tasa de aprendizaje (learning rate)	2e-5

Al igual que los hiperparámetros, es importante definir el optimizador y la función de pérdida, para todos los casos se hace uso de AdamW[4], que es un optimizador mejorado de Adam, produce una mejor pérdida de entrenamiento y recibe como parámetro la tasa de aprendizaje (learning rate) que es ajustada con Scheduler[5], el programador de tasas de aprendizaje que ajustará el modelo para encontrar la tasa óptima en caso de que la estipulada no lo fuera. En el caso de la función de pérdida se calcula mediante la función CrossEntropyLoss que recibe como parámetro los pesos calculados como vector que se aplica a cada clase, como se mencionó anteriormente.

G. *Entrenamiento y prueba*

Definido todo lo requerido, se procede con el entrenamiento de los modelos haciendo uso del método <model>.train() y las validaciones con <model>.eval().

Para evaluar los resultados se hizo uso de métricas como accuracy que mide la cantidad de aciertos en comparación de la cantidad de datos totales evaluados, precision mide los resultados de predicción realmente correctos, recall (sensibilidad) indica la capacidad del modelo en clasificar correctamente una clase y F1-Score es la combinación entre recall y precision. También se generó la matriz de confusión que muestra las predicciones correctas e incorrectas obtenidas en comparación con las etiquetas reales.

III. RESULTADOS

Mediante la métrica accuracy se analizaron los valores obtenidos para determinar los resultados de los modelos entrenados. La evaluación de los modelos con el dataset de prueba para comparar el desempeño de los modelos en la tarea de predicción se realizó con las métricas de precision, recall y F1-score para cada clase. Tabla 6 muestra los resultados para los modelos en el entrenamiento con noticias en inglés y Tabla 7 para las de español.

TABLA 6
EVALUACIÓN DE MODELOS EN INGLÉS

Modelo	Accuracy	Categoría	Precision	Recall	F1-Score	Tiempo
DistilBerT-base-uncased	0,39	Sátira	0,86	0,12	0,21	6 min, 4 s
		Opinión	0,23	0,94	0,38	
		Información	0,92	0,61	0,73	
RoBERTa-base	**0,73**	Sátira	0,00	0,00	0,00	12 min, 36 s
		Opinión	0,77	0,91	0,83	
		Información	1,00	0,41	0,59	

TABLA 7
EVALUACIÓN DE MODELOS EN ESPAÑOL

Modelo	Accuracy	Categoría	Precision	Recall	F1-Score	Tiempo
BERT-base-spanish-wwm-cased	0,59	Sátira	0,71	0,16	0,26	16 min, 16 s
		Opinión	0,53	0,92	0,67	
		Información	0,83	0,67	0,74	
RoBERTa-base-BNE	0,37	Sátira	0,86	0,11	0,20	16 min, 3 s
		Opinión	0,30	0,90	0,45	
		Información	0,50	0,67	0,57	
IXAmBERT-base-cased	0,72	Sátira	0,14	0,12	0,13	15 min, 13 s
		Opinión	0,81	0,84	0,83	
		Información	0,58	0,54	0,56	
XLM-RoBERTa-base	**0,77**	Sátira	0,00	0,00	0,00	15 min, 54 s
		Opinión	1,00	0,77	0,87	
		Información	0,00	0,00	0,00	
mDeBERTa-v3-base	0,70	Sátira	0,00	0,00	0,00	23 min, 0 s
		Opinión	0,91	0,75	0,82	
		Información	0,00	0,00	0,00	

Para inglés, DistilBERT-base-uncased logró un accuracy del 39%, mientras que con RoBERTa-base se obtuvo 73%. Para español, BERT-case-spanish-wwm-cased obtuvo un accuracy de 59%, mientras que XLM-RoBERTa-large alcanzó los 37%, en cambio, IXAmBERT-base-cased logró 72%, XLM-RoBERTa-base 77%, por último, mDeBERTa-v3-base consiguió 70%.

En la matriz de confusión podemos observar el comportamiento de las predicciones de los modelos con el conjunto de datos de pruebas. Fig 5 y Fig 6 para análisis de noticias en inglés y español respectivamente.

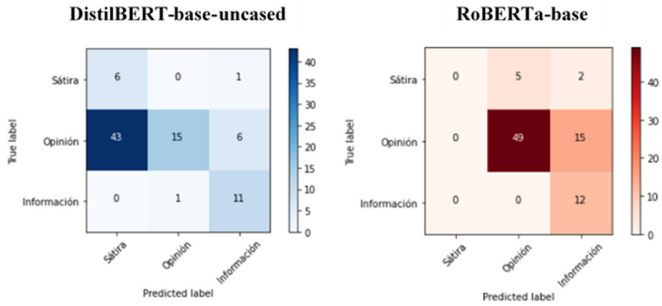

Fig. 5 Matriz de confusión obtenida de modelos en inglés.

[4] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html

[5] https://pytorch.org/docs/stable/optim.html

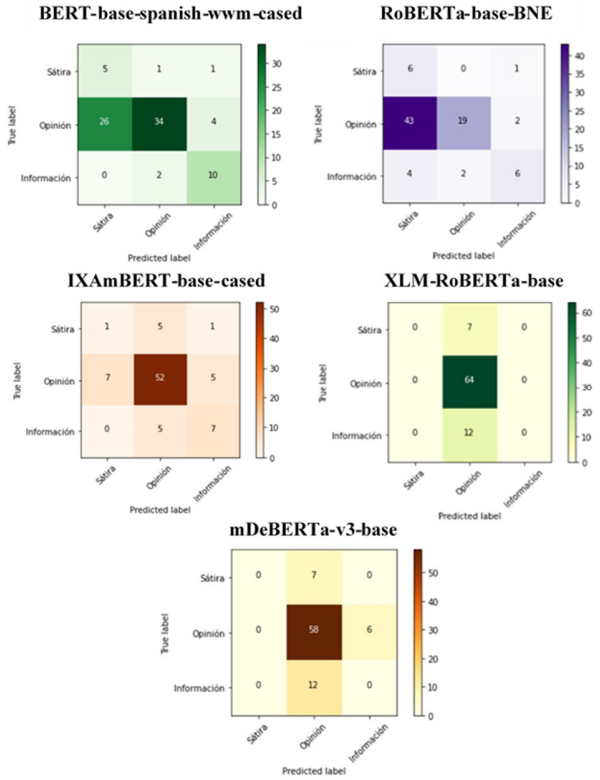

Fig. 6 Matriz de confusión obtenida de modelos en español.

Con las métricas producidas se observa el rendimiento general de los modelos, tanto para el idioma inglés como para el español.

IV. Discusión

Se evidencia el modelo RoBERTa-base mejor en la categorización de noticias en inglés con un accuracy del 73%. DistilBERT-base-uncased logró buena precisión en los artículos de sátira (0,86), aunque el recall es bajo (0,12), lo que implica que no todas las que predijo como sátira eran realmente de ese tipo. En cambio, para los artículos de opinión, si bien tiene alto recall (0,94) que significa que la gran mayoría que predijo como opinión eran exactamente de ese tipo, no fue capaz de predecir todos los artículos de opinión. En los artículos de información fue más equilibrado logrando predecir casi todas las muestras (0,92), si bien algunas que predijo como información no eran así. De igual manera, RoBERTa-base obtuvo mejores resultados para los artículos de opinión e información donde la precisión fue mayor (0,77 y 1,00 respectivamente), aunque este no fue capaz de identificar las sátiras, que al parecer fueron predichas como tipo información.

Para los modelos en español, XLM-RoBERTa-base obtuvo el mayor accuracy de 77% pero solo predice opinión, de igual manera para mDeBERTa-v3-base que a pesar de tener accuracy de 70% solo predice opinión. Esto puede deberse a que los modelos se entrenaron con un dataset en el que había muchas más noticias de opinión que las demás y al parecer la asignación de pesos para cada clase con CrossEntropyLoss no fue suficiente para tratar el desbalance. En cambio, IXAmBERT-base-cased con 72% de accuracy predice de las tres categorías. La sátira es una categoría subjetiva, lo que causa que obtenga un bajo resultado de predicción (0,14). Los modelos con más bajo accuracy fueron BERT-base-spanish-wwm-cased con 59% y RoBERTa-base-BNE con 37% que en ambos casos tienen recall alto en las noticias de opinión (0,92 y 0,90 respectivamente), lo que demuestra que la mayor parte de noticias que predijeron como opinión realmente lo eran, aunque se observa precision bajo de ambos modelos, sobre todo en RoBERTa-base-BNE, lo que significa que no predijo todas las muestras, ya que al parecer estas fueron tomadas como las otras categorías, principalmente como sátira, esto se ve reflejado en el recall de la clase para ambos modelos.

En general, con las métricas producidas se observa que los modelos se desenvolvieron mejor con los artículos de opinión. Sátira y opinión son noticias de tipo subjetivo, es probable que estos sistemas sean más capaces de enmarcarlas mejor las noticias de opinión y tomar las de sátira como opinión. Las noticias de tipo información fueron la segunda categoría con mejor rendimiento, estas son de tipos objetivas, por lo que son más definidas para los sistemas.

El conjunto de datos original es de noticias en idioma inglés, por motivos de experimentación se tradujeron al idioma español utilizando DeepL Translator[6] para realizar pruebas también en este idioma. Es probable que esta haya sido una de las razones por la que, en algunos modelos de clasificación para el idioma español como mDeBERTa-v3-base y XLM-RoBERTa-base, no hayan podido clasificar las noticias de tipo satíricas. Existe una pérdida en la calidad de la traducción del inglés al español sobre todo con las noticias de tipo satírico que estos sistemas no los pueden hacer de forma plana ni retener su sentido original. Por otro lado, con las noticias de opinión, estos modelos de predicción fueron capaces de enmarcarlas mejor con el dataset traducido.

Las noticias de información al ser objetivas son más definidas para los sistemas, por este motivo, en todos los modelos, tanto para inglés como español, la métrica de F1 score fue más alta.

Otra de las razones por las que puede deberse la precisión baja en las categorías de sátira e información, es el tamaño del dataset usado para el entrenamiento, ya que solo contenía 433 noticias. Además, presentaba un desbalance en las clases, siendo la categoría de opinión la que más muestras contenía, a pesar de utilizar la técnica de pérdida de entropía cruzada para tratar este desbalance, no resultó de gran ayuda en algunos modelos, siendo más evidente en el modelo para inglés RoBERTa-base en donde, las noticias de tipo sátira, el sistema no le prestó la atención necesaria, resultando como cero en las

[6] https://www.deepl.com/translator

métricas aplicadas. Asimismo, en los modelos para español mDeBERTa-v3-base y XLM-RoBERTa-base, en donde prevaleció la clase dominante opinión, como resultado generalizó todas las noticias como opinión, a excepción de mDeBERTa-v3-base que hubo seis noticias predichas como información, pero que realmente eran de opinión.

V. Conclusiones

Para el idioma inglés el modelo RoBERTa-base resultó superior en la clasificación de las categorías de opinión e información, sin embargo, no reconoció las de sátira. DistilBERT pudo predecir algunos de las muestras, aunque no resultó ser tan fiable sobre todo en las sátiras.

Para el español, XLM-RoBERTa-base obtuvo el mayor accuracy pero no consideró las noticias de tipo sátira e información, algo similar pasó con mDeBERTa-v3-base que tampoco fue capaz de reconocerlas. Al que le fue peor en la predicción de sátira fue IXAmBERT-base-cased y el mejor fue BERT-base-spanish-wwm-cased., este último también resultó superior en identificar las de información. Los modelos más equilibrados en cuanto a predicción de todas las categorías establecidas en esta investigación para noticias en español fueron IXAmBERT-base-cased y BERT-base-spanish-wwm-cased.

Los resultados de predicción obtenidos demostraron que mediante el ajuste fino de modelos Transformer pre entrenados se puede lograr categorizar noticias en diversos idiomas en específico para determinar si son sátira, opinión o información.

Para trabajos futuros relacionados a la clasificación de textos en español, se recomienda que el dataset sea de noticias provenientes de fuentes en español y, de ser posible, presente un mejor balance de las clases, también se puede implementar otras técnicas de balanceo. Además, para las noticias de tipo sátira y opinión que son de tipo subjetivo, se podría aplicar métodos basados en el análisis estilométrico como longitud de palabras, de frase, tipo de escritura, etc., en combinación con los modelos Transformer.

V. Referencias

[1] P. Doblas Martin, "News classification using natural language processing techniques based on deep learning Cotutorizado por," 2021.

[2] A. Onan and M. A. Tocoglu, "Satire identification in Turkish news articles based on ensemble of classifiers," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 2, pp. 1086–1106, Jan. 2020, doi: 10.3906/elk-1907-11.

[3] C. Jashubhai Rameshbhai and J. Paulose, "Opinion mining on newspaper headlines using SVM and NLP," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 3, pp. 2152–2163, 2019, doi: 10.11591/ijece.v9i3.pp2152-2163.

[4] L. M. Gutiérrez-Coba, P. Coba-Gutiérrez, and J. A. Gómez-Díaz, "Fake news about Covid-19: A comparative analysis of six iberoamerican countries," *Revista Latina de Comunicacion Social*, vol. 2020, no. 78, pp. 237–264, 2020, doi: 10.4185/RLCS-2020-1476.

[5] L. Talamé, A. Cardoso, and M. Amor, "Comparación de herramientas de procesamiento de textos en español extraídos de una red social para Python," *Simposio Argentino de Inteligencia Artificial*, vol. 1, pp. 53–67, 2019, Accessed: Jan. 11, 2023. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/87854

[6] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved Bi-LSTM-CNN," *Proceedings - 9th International Conference on Information Technology in Medicine and Education, ITME 2018*, pp. 890–893, 2018, doi: 10.1109/ITME.2018.00199.

[7] P. K. Mallick, S. Mishra, and G. S. Chae, "Digital media news categorization using Bernoulli document model for web content convergence," *Pers Ubiquitous Comput*, 2020, doi: 10.1007/s00779-020-01461-9.

[8] T. B. Shahi and A. K. Pant, "Nepali news classification using Naïve Bayes, Support Vector Machines and Neural Networks," in *2018 International Conference on Communication information and Computing Technology (ICCICT)*, IEEE, Feb. 2018, pp. 1–5. doi: 10.1109/ICCICT.2018.8325883.

[9] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, no. 1, 2020, doi: 10.1007/s41133-020-00032-0.

[10] I. N. Santana, R. S. Oliveira, and E. G. S. Nascimento, "Text Classification of News Using Transformer-based Models for Portuguese," *J Syst Cybern Inf*, vol. 20, no. 5, pp. 33–59, 2022, doi: 10.54808/jsci.20.05.33.

[11] T. Chakraborty, K. Shu, · H Russell Bernard, H. Liu, · Md, and S. Akhtar, "Revised Selected Papers Communications in Computer and Information Science 1402," 2021. [Online]. Available: http://www.springer.com/series/7899

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, doi: 10.48550/arxiv.1810.04805.

[13] R. Singh, S. A. Chun, and V. Atluri, "Developing machine learning models to automate news classification," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jun. 2020, pp. 354–355. doi: 10.1145/3396956.3397001.

[14] J. F. Low, B. C. M. Fung, F. Iqbal, and S. C. Huang, "Distinguishing between fake news and satire with transformers," *Expert Syst Appl*, vol. 187, p. 115824, Jan. 2022, doi: 10.1016/J.ESWA.2021.115824.

[15] P. Saigal and V. Khanna, "Multi-category news classification using Support Vector Machine based classifiers," *SN Appl Sci*, vol. 2, no. 3, pp. 1–12, Mar. 2020, doi: 10.1007/S42452-020-2266-6/TABLES/9.

[16] C. H. Espin-Riofrio, K. I. Vera-Guamán, and R. Yela-García III, "Clasificación y etiquetado de tweets de Ecuador para determinar qué tema tratan, utilizando un modelo Transformer," *Polo del Conocimiento: Revista científico - profesional, ISSN-e 2550-682X, Vol. 7, Nº. 3, 2022*, vol. 7, no. 3, p. 34, 2022, doi: 10.23857/pc.v7i3.3791.

[17] J. Ortiz-Zambrano, C. Espin-Riofrio, and A. Montejo-Ráez, "Transformers for Lexical Complexity Prediction in Spanish Language," *Procesamiento del Lenguaje Natural*, vol. 69, pp. 177–188, 2022, doi: 10.26342/2022-69-15.

[18] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, "Machine Learning for Fluid Mechanics," *https://doi.org/10.1146/annurev-fluid-010719-060214*, vol. 52, pp. 477–508, Jan. 2020, doi: 10.1146/ANNUREV-FLUID-010719-060214.

[19] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What is Machine Learning? A Primer for the Epidemiologist," *Am J Epidemiol*, vol. 188, no. 12, pp. 2222–2239, Dec. 2019, doi: 10.1093/AJE/KWZ189.

[20] A. López Díaz, "Fundamentos matemáticos de los métodos Kernel para aprendizaje supervisado," 2018, Accessed: Jan. 27, 2023. [Online]. Available: https://idus.us.es/handle/11441/77547

[21] V. L. A. Chamorro, "Clasificación de tweets mediante modelos de aprendizaje supervisado," 2018.

[22] K. Bajaj and G. Kaur, "News Classification using Neural Networks News Classification and Its Techniques: A Review," *Journal of Computer Engineering*, vol. 18, no. 1, pp. 22–26, 2016, doi: 10.5120/cae2016652224.

[23] L. J. Sandoval Serrano, "Algoritmos de aprendizaje automático para análisis y predicción de datos," vol. 11, Oct. 2018, Accessed: Jan. 27, 2023. [Online]. Available: http://redicces.org.sv/jspui/handle/10972/3626

[24] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[25] A. Vaswani *et al.*, "Attention is All you Need," *Adv Neural Inf Process Syst*, vol. 30, 2017.

[26] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 328–339, Jan. 2018, doi: 10.48550/arxiv.1801.06146.

[27] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018. [Online]. Available: https://gluebenchmark.com/leaderboard

[28] "gpt2 · Hugging Face." https://huggingface.co/gpt2 (accessed Jan. 17, 2023).

[29] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Adv Neural Inf Process Syst*, vol. 2020-December, May 2020, doi: 10.48550/arxiv.2005.14165.

[30] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. v. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 2978–2988, Jan. 2019, doi: 10.48550/arxiv.1901.02860.

[31] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. v. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *Adv Neural Inf Process Syst*, vol. 32, 2019, Accessed: Jan. 23, 2023. [Online]. Available: https://github.com/zihangdai/xlnet

[32] Hugging Face, "Pytorch-Transformers — documentación de pytorch-transformers 1.0.0." https://huggingface.co/transformers/v1.2.0/ (accessed Jan. 23, 2023).

[33] M. Honnibal and I. Montani, "spaCy meets Transformers: Fine-tune BERT, XLNet and GPT-2 · Explosion," 2019. https://explosion.ai/blog/spacy-transformers (accessed Jan. 23, 2023).

[34] A. Conneau and G. Lample, "Cross-lingual Language Model Pretraining," *Adv Neural Inf Process Syst*, vol. 32, Jan. 2019, doi: 10.48550/arxiv.1901.07291.

[35] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," pp. 7871–7880, Oct. 2019, doi: 10.48550/arxiv.1910.13461.

[36] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020, Accessed: Jan. 24, 2023. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[37] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019, doi: 10.48550/arxiv.1910.01108.

[38] A. Hande, K. Puranik, R. Priyadharshini, S. Thavareesan, and B. R. Chakravarthi, "Evaluating Pretrained Transformer-based Models for COVID-19 Fake News Detection," *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, pp. 766–772, Apr. 2021, doi: 10.1109/ICCMC51019.2021.9418446.

[39] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, doi: 10.48550/arxiv.1907.11692.

[40] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA," 2020. [Online]. Available: https://github.com/josecannete/spanish-corpora

[41] A. Otegi, A. Agirre, J. A. Campos, A. Soroa, and E. Agirre, "Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque," pp. 11–16, 2020, Accessed: Feb. 03, 2023. [Online]. Available: http://ixa.si.ehu.es/node/12934

[42] A. Gutiérrez-Fandiño *et al.*, "MarIA: Spanish Language Models MarIA: Modelos del Lenguaje en Español," pp. 39–60, 2022, doi: 10.26342/2022-68-3.

[43] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," Nov. 2019, doi: 10.48550/arxiv.1911.02116.

[44] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," Nov. 2021, Accessed: Feb. 03, 2023. [Online]. Available: http://arxiv.org/abs/2111.09543

[45] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," Jun. 2020, Accessed: Feb. 20, 2023. [Online]. Available: http://arxiv.org/abs/2006.03654