applied sciences

MDPI

Article

Enhancing Misinformation Detection in Spanish Language with Deep Learning: BERT and RoBERTa Transformer Models

Yolanda Blanco-Fernández [1,*], Javier Otero-Vizoso [2], Alberto Gil-Solla [1] and Jorge García-Duque [2]

[1] atlanTTic Research Center for Telecommunication Technologies, University of Vigo, 36310 Vigo, Spain; agil@det.uvigo.es
[2] Escuela de Ingeniería de Telecomunicación, University of Vigo, 36310 Vigo, Spain; jgd@det.uvigo.es (J.G.-D.)
* Correspondence: yolanda@det.uvigo.es

Abstract: This paper presents an approach to identifying political fake news in Spanish using Transformer architectures. Current methodologies often overlook political news due to the lack of quality datasets, especially in Spanish. To address this, we created a synthetic dataset of 57,231 Spanish political news articles, gathered via automated web scraping and enhanced with generative large language models. This dataset is used for fine-tuning and benchmarking Transformer models like BERT and RoBERTa for fake news detection. Our fine-tuned models showed outstanding performance on this dataset, with accuracy ranging from 97.4% to 98.6%. However, testing with a smaller, independent hand-curated dataset, including statements from political leaders during Spain's July 2023 electoral debates, revealed a performance drop to 71%. Although this suggests that the model needs additional refinements to handle the complexity and variability of real-world political discourse, achieving over 70% accuracy seems a promising result in the under-explored domain of Spanish political fake news detection.

Keywords: fake news; Spanish; curated synthetic dataset; fine-tuning; Transformer-based models; BERT; RoBERTa

check for updates

Citation: Blanco-Fernández, Y.; Otero-Vizoso, J.; Gil-Solla, A.; García-Duque, J. Enhancing Misinformation Detection in Spanish Language with Deep Learning: BERT and RoBERTa Transformer Models. *Appl. Sci.* **2024**, *14*, 9729. https://doi.org/10.3390/app14219729

Academic Editor: Douglas O'Shaughnessy

Received: 13 September 2024
Revised: 3 October 2024
Accepted: 15 October 2024
Published: 24 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the current information age, where data flow rapidly, the battle against fake news is crucial to maintain quality. Misinformation, particularly in political contexts, threatens public trust in the media and politics, affecting vital areas such as the economy, healthcare, and education. The need to detect and counteract misinformation has led to advances in Machine Learning (ML) for fake news detection. However, most research has focused on general news in English, leaving a significant gap in detecting political misinformation in Spanish [1–5]. Recent breakthroughs in Deep Learning (DL), particularly with Transformer models like BERT [6], have demonstrated exceptional performance in natural language processing (NLP) tasks, including text classification for fake news detection [7–10].

Transformer models in NLP tasks operate through two key phases: pre-training and fine-tuning. During pre-training, the model is exposed to large amounts of unlabeled text, building a foundational understanding of language through techniques like word masking and Next Sentence Prediction (NSP). This phase enables broad language comprehension. Fine-tuning, on the other hand, adapts the model for specific tasks using a smaller, labeled dataset. Pre-training imparts general linguistic knowledge, while fine-tuning refines the model for task-specific performance, making it highly effective for various NLP tasks.

Transformer models, with their attention mechanism [11], are highly effective at understanding context and word relationships, making them well suited for detecting subtle patterns in fake news. Their ability to grasp nuances is critical for tasks like fake news detection. However, a major challenge in identifying political fake news in Spanish is the limited availability of robust datasets. Existing Spanish datasets, such as those from Kaggle, are often small and fail to cover the broad range of political issues, including

economic, healthcare, and education matters [12–14]. This lack of comprehensive data continues to be a significant obstacle in political fake news detection in Spanish.

This research presents a comprehensive solution to address these challenges. We introduce an approach used to autonomously generate a dataset of 57,231 political news articles in Spanish (46,000 used for training and 11,231 for testing). Many of these news articles are gathered from reputable sources through web scraping and are also generated using information released by Spain's National Institute of Statistics (INE) (the acronym INE refers to the Instituto Nacional de Estadística in Spanish), an authoritative body responsible for the collection, analysis, and distribution of official statistical data within the country. Our dataset greatly benefits from the availability of reliable information provided by the INE, encompassing demographic metrics, economic indicators, and statistics related to health and education. These data serve as both a robust source and an essential context for the detection of political fake news in the Spanish language. Due to the difficulties in finding trustworthy fake news samples, the dataset comprises around 28,000 fake news articles acquired through two methods: (i) employing generative models developed by OpenAI and (ii) creating fake news articles by automatically modifying specific details in real news, such as dates, statistics, and the identities of political figures, using Python libraries and scripts to falsify their content.

The main contribution of our paper can be summarized as follows:

- Creation of an extensive dataset of Spanish political news, automatically collected from reliable sources and enriched with generative artificial intelligence techniques.
- Fine-tuning and evaluation of Transformer models, including BERT and its optimizations, for political fake news detection.
- Comparison of our results with those from related works, highlighting the performance of our fine-tuned BERT and RoBERTa models for Spanish political content, a theme and language that are underexplored in the literature.

This paper is structured as follows: In Section 2, we present an overview of related work in the domain of fake news detection. Section 3 evaluates the strengths and weaknesses of the current training datasets utilized in Spanish fake news detection and details the establishment of our dedicated dataset for this research, aimed at addressing identified gaps and limitations. In Section 4, we provide a comprehensive overview of BERT-based models, emphasizing optimizations such as RoBERTa, AlBERT, and DistilBERT, and highlight the fine-tuning process with our Spanish political fake news dataset. Section 5 features an experimental discussion, encompassing performance metrics for the resulting fine-tuned models, along with a benchmark comparison against prior research. Finally, Section 6 concludes the paper and outlines potential lines for further research.

2. Related Work

Transformer-based models fine-tuned on existing corpora have also been employed for generic NLP text classification tasks with diverse purposes, beyond the particular fake news detection tackled in our research, which can essentially be seen as a categorization of textual input as fake or true. In this domain, we find works like [15], which aims to identify sexist content on social media using a multilingual BERT model to classify tweets as either sexist or non-sexist, employing a dataset of 13,000 texts (half in English and half in Spanish) to achieve an accuracy of 76.83%.

Similarly, the properties and performance of fine-tuned BERT models are also explored in [16] for classification tasks in different languages and purposes, including sentiment analysis of IMDB movie reviews (50,000 entries in English, accuracy of 93.87%), hotel reviews in Chinese (6000 reviews, accuracy of 93.81%), classification of informative tweets related to natural disasters (10,000 tweets in English, accuracy of 83.64%), and categorization of news articles in Portuguese across multiple content categories (12,145 headlines, accuracy of 91.20%).

Moving already to the particular domain of fake news detection, the identification of such news has been tackled by multiple conventional ML techniques, including Random

Forest (RF), Decision Trees (DT), Boosting Trees (BT), logistic regression (LR), Support Vector Machine (SVM), and Naïve Bayes (NB) [17]. These techniques, while effective in various contexts, exhibit several limitations when compared to DL models that have achieved better performance in fake news detection tasks. In particular, one key limitation of traditional ML models like NB and SVB is their reliance on manual feature engineering. In these approaches, human experts must define and select features such as word frequencies, n-grams, or specific keywords to capture patterns in the data. This process can be time-consuming and may fail to capture complex patterns or deeper contextual information. Additionally, traditional models like LR often assume linear relationships between the selected features and the target variable, which limits their ability to handle more intricate interactions within the data. These models also tend to fail to generalize across diverse datasets, particularly when there are variations in linguistic or stylistic patterns. In contrast, DL models have the advantage of automatically learning features from raw data, capturing both simple and complex patterns without manual intervention. By utilizing multiple layers of representation, DL approaches can more effectively model intricate relationships and contextual dependencies in large datasets, making them far more adaptable and powerful for tasks like fake news detection [18–20].

Given the aforementioned weaknesses, traditional ML techniques have sometimes been integrated with DL models to leverage the strengths of both approaches. Specifically, ML methods have been combined with Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) [21–24], and also with various types of word and document embeddings and Transformer-based models, resulting in notable outcomes [25–27]. For instance, the study presented in [28] introduces an effective approach for detecting fake news on social media by utilizing three publicly available datasets (WELFake, FakeNewsNet, and FakeNewsPrediction) and integrating FastText word embeddings with various ML and DL techniques. Specifically, a hybrid model of CNN and LSTM achieved high accuracy, while Transformer models like BERT, XLNet, and RoBERTa showed better performance after hyperparameter tuning, achieving accuracy values of 99%. Additionally, explainable artificial intelligence (AI) methods were employed to elucidate the model's decision-making process.

The detection of fake news in social media has been tackled using various methods, including graph convolutional networks and attention networks. The former leverages the structural relationships in graph data to analyze connections and patterns within the social network, aiding in identifying misleading information [29]. Meanwhile, attention networks focus on important parts of the input data by assigning different levels of importance, allowing the model to prioritize relevant features, better detect nuances indicative of fake news, and even manage handle multimodal misinformation detection in multimedia content [30].

Most studies have predominantly concentrated on English-language news, showcasing notable outcomes with high accuracy rates within the range of 84% to 92% [3,31–33]. In [12], detection is performed on a relatively small dataset containing 2000 news articles in Spanish, equally divided between real and fake news, which was created in [34]. As depicted in Table 1, the obtained accuracy values are as follows: 64.25% with XGBoost, 77% with logistic regression, and 74.25% with Random Forest. In [14], various classic ML methods are explored such as DT, SVM, NB and K-Nearest Neighbors (KNN), for classifying Mexican Spanish news as either 'fake' or 'true'. This experiment employs a news corpus comprising pairs of real and fake Spanish-language publications collected between November 2020 and March 2021 from newspaper websites, media companies, and fact-checking platforms. The dataset consists of 971 instances, with balanced representation of fake and true news in the test dataset [35]. Experimental results reveal that SVM and KNN are the most effective classification models in this context, achieving an accuracy of 81%.

Table 1. Related works about fake news detection and text classification by ML and Transformer models.

Related Work	Adopted Technique	Dataset	Size	Language	Accuracy
[28]	BERT, XLNet, RoBERTa	WELFake, FakeNewsNet, and FakeNewsPrediction	72,134	English	99%
[15]	BERT	EXIST [36]	13,000	English Spanish	76.83%
[16]	BERT	IMDB movie reviews [37] RealOrNot tweets [38] Portuguese news [39] Chinese hotel reviews [40]	50,000 10,000 12,145 6000	English English Portuguese Chinese	93.87% 83.64% 91.20% 93.81%
[12]	Random Forest XGBoost Logistic Regression	Fakes1000 [34]	2000	Spanish	74.25% 64.25% 77%
[14]	SVM Naive Bayes KNN (K = 3) Decision Tree	The Spanish Fake News corpus [35]	971	Mexican Spanish	81% 59% 81% 62%
[13]	Random Forest BERT	Unpublished dataset	634	Spanish	89.80% 95.80%
[33]	Two parallel BERT networks	Fakenewsnet [41]	23,000	English	85.4%
[42]	BETO	Corpus extracted from [34,43]	2500	Spanish	80%

In [13] an approach to detect fake news in Spanish, specifically in tweets posted during the 2022 Colombian presidential elections, is described. This work compares the performance of ML and DL techniques by pitting RF against a fine-tuned BERT model. In both cases, a relatively small corpus comprising just over 600 Spanish tweets related to the 2022 elections, equally divided between fake and real news, was used, obtaining an accuracy of 89.90% with RF and 95.80% with BERT.

The superior performance of Transformer models in text classification tasks has been confirmed in various research studies using BERT models and variants for fake news detection in quite different languages, such as English [23], Greek [44], and Bengali [45]. These studies achieved accuracy values of 83%, 93%, and 92%, respectively. The authors of [33] proposed MVPBert, which utilizes two parallel BERT networks to perform veracity detection in English using the Fakenewsnet dataset containing approximately 23,000 news items [41]. Fakenewsnet is a significant dataset consisting of full-length news articles obtained from https://www.politifact.com (accessed on 1 November 2023). In contrast to other fake news datasets that often comprise shorter sources like tweets and headlines, Fakenewsnet provides extensive news articles. In this work, one BERT network encodes news headlines, while the other encodes the news body. Due to the constant and limited input length of the BERT network and the typically longer news body, the authors employ the MaxWorth algorithm to select the most valuable part of the news text for fact checking, achieving an accuracy value of 85.4%.

The approach outlined in [42] employs a BETO model, a dedicated architecture trained from scratch with documents in Spanish. In particular, this model, designed as the Spanish counterpart of BERT (originally developed for English), has undergone fine-tuning using a corpus containing 2500 Spanish news articles. The achieved accuracy rate exceeds 80% [34,43].

The revision of existing approaches made throughout this section reveals that Transformer models, notably BERT and its variants, outperform traditional ML methods for fake news detection, achieving higher accuracy rates and robustness across various languages and contexts. However, when focusing on detecting fake news in Spanish, especially in the political context, a significant challenge arises due to the lack of comprehensive Spanish training datasets. Our paper fills this gap by assembling an extensive and varied collection of high-quality Spanish political news articles, as we will describe in Section 3. This dataset has been curated specifically for the purpose of fine-tuning and evaluating Transformer-

based models designed to detect fake news in Spanish within the domain of politics, a realm that has been relatively underexplored in the current state-of-the-art methodologies. As it will be detailed in Section 4, the objective is to illustrate the performance of these models, which serve as valuable instruments in the fight against misinformation in Spanish-speaking areas.

3. Crafting a Spanish Fake News Dataset

In this section, we will discuss the main shortcomings identified in current Spanish-language datasets and provide detailed insights into the criteria considered in the development of our customized corpus (Section 3.1). Additionally, we will detail the procedure, sources, and tools utilized for the automated collection of our Spanish political fake news dataset (Section 3.2).

3.1. Known Limitations and Desired Features

The need to establish a Spanish-language dataset dedicated to political disinformation arises from the inherent constraints identified in current datasets, as outlined in Section 2. These limitations have impeded the development of robust and accurate classification models, which are essential for combating disinformation within the Spanish-language political context. The identified constraints involve the following features:

- Size of the datasets. Current Spanish datasets are notably small, with a total number of news articles tagged as 'fake' or 'real' that falls below 3000 [12–14]. This lack of data compromises the ability of classification models to effectively learn to distinguish between true and false news.
- Topic of the news. These datasets encompass a broad spectrum of topics extending beyond the Spanish political context, making it challenging to build models specialized in the detection of political fake news.
- Format errors in news. Errors have been observed in the news articles, such as missing fields and incorrect categorizations, affecting the quality of the data. Additionally, current fake news datasets often lack temporal references, potentially leading to inaccurate and biased results in classification models since such references are indispensable for evaluating the credibility of news content. Significantly, data such as the current president's name or statistics regarding unemployment and economic growth are inherently time sensitive. It is imperative that news articles within these datasets incorporate temporal markers to guarantee the seamless functioning of fake news detection models, ultimately improving the accuracy of misinformation identification.

To overcome these limitations, this paper describes the creation of a customized dataset with the following requirements:

- Our dataset should contain a significantly larger volume of news articles to address the issue of small size.
- Furthermore, the new corpus should collect genuine news from reputable and trustworthy sources. In the absence of reliable equivalent sources for fake news, a procedure for generating fake news in a balanced manner becomes necessary. This should ensure a comprehensive and equitable dataset that reflects both real and fake information, thereby enhancing the model's ability to accurately distinguish between them and, consequently, address misinformation effectively.
- The new dataset should include news articles of varying lengths, allowing the model to process headlines of different extensions without compromising its effectiveness, while also addressing issues related to formatting errors in the news articles.
- Regarding thematic diversity, the news articles in the new dataset should cover a variety of topics within the political context, such as economy, society, culture, among others. This will enable the model to verify information considering the vocabulary used in the articles, thus addressing the issue of excessively broad subject matter in current datasets.

- Finally, it is crucial to ensure that the new dataset contains both subtle and significant differences between real and fake news articles. This is relevant because, as the similarity in language and structure between real and fake news intensifies, the task of differentiation becomes more challenging for the model. Therefore, the dataset creation process should meticulously highlight these nuanced distinctions. This approach empowers the model to more effectively identify even the most subtle variations between accurate information and misleading content, thereby enhancing its proficiency in combating misinformation.

3.2. Developing a Custom Dataset for Spanish Politics-Related Fake News

As shown in the brief excerpt from our Spanish Political Fake News dataset depicted in Figure 1, this corpus includes the following fields for each news article:

- ID: A unique identifier for each news item.
- Label: A classification label designating each news item as either fake (0) or real (1).
- Headline (*Titulo* in Figure 1): The headline of the news article.
- Description (*Descripcion* in Figure 1): A more detailed description of the news article, providing additional insights into the subject matter, personalities, or events mentioned within.
- Date (Fecha in Figure 1): The publication date of the news articles in DD/MM/YYYY format, commonly adopted in Spanish.

	ID	Label	Titulo	Descripcion	Fecha
0	1	1	Moreno intenta apaciguar el flanco sanitario m...	El presidente abre la puerta a unos comicios e...	19/04/2022
1	2	1	La Abogacía del Estado se retira como acusació...	En un escrito, la abogada del Estado Rosa Marí...	17/09/2021
2	3	0	Las promesas incumplidas de Pablo Echenique en...	Este lunes y martes la Asamblea de Madrid acog...	12/09/2022
3	4	1	Sánchez defiende 'resolver el problema' de la ...	Resulta evidente que la ley ha tenido algunos ...	07/02/2023
4	5	1	Ian Gibson cierra la lista electoral de la con...	El hispanista, que ya ocupó un puesto simbólic...	12/04/2023
5	6	1	ERC y Junts se atascan en la negociación para ...	Empieza una intensa semana para el Ejecutivo c...	19/09/2022
6	7	1	El PP se rinde al discurso antiabortista de Vo...	Núñez Feijóo guarda silencio pero Ayuso y Alme...	13/01/2023
7	8	1	El PSOE denuncia que el hombre de Mañueco en S...	Los socialistas han remarcado que estos trabaj...	31/03/2023
8	9	1	Escrivá niega que su reforma de Extranjería ge...	El ministro de Inclusión, Seguridad Social y M...	08/08/2022
9	10	1	IU propone una descentralización para empodera...	Izquierda Unida debate y aprueba este sábado s...	26/11/2022
...					
	ID	Label	Titulo	Descripcion	Fecha
21442	21443	1	Amnistía Internacional pide al Gobierno españo...	La organización de derechos humanos alerta sob...	29/07/2022
21443	21444	1	La comunidad educativa, contra la eliminación ...	El PSOE, saltándose el acuerdo que suscribió, ...	17/11/2021
21444	21445	1	La campaña de Hazte Oír contra Casado y a favo...	El lobby ultraconservador aprovecha la grave c...	21/02/2022
21445	21446	0	Rosalía Lluís Rabell, esposa de Pío García Esc...	Ha declarado en la comisión del Congreso de lo...	13/05/2021
21446	21447	1	El ex coordinador de los antidisturbios de Mad...	El inspector jefe de Policía, ya jubilado, acu...	12/04/2023
21447	21448	1	La vacunación a la Policía Nacional y la Guard...	El objetivo es alcanzar el jueves próximo el 8...	03/05/2021
21448	21449	0	La fundación ultrapatriota de Vox dejó de publ...	DENAES, creada con el objetivo de la defensa d...	06/02/2019
21449	21450	1	Ione Belarra presenta su candidatura a la Secr...	La ministra y dirigente de la formación morada...	12/05/2021
21450	21451	1	Casado y Ayuso condenan las amenazas a Iglesia...	No condenar este tipo de agresiones personales...	23/04/2021
21451	21452	0	Escándalo político en el Congreso por presunta...	Un escándalo político ha sacudido al Congreso ...	12/02/2022

Figure 1. Excerpt from our dataset of news articles related to politics in the Spanish language, which includes a total of 57,231 entries, comprising 52% real news and 48% fake news.

The process of constructing our dataset of fake news in Spanish involves a combination of different techniques, which we will explore throughout this section.

- Firstly, we employed web scraping to automatically extract real news from information published between April 2017 and June 2023. These data were gathered from the websites of reputable, independent media outlets that have established a strong reputation for reliability and impartiality. These outlets are known for avoiding political bias in their coverage of a wide range of topics, specifically those related to national politics. This process allowed us to collect authentic news articles that provide a balanced and accurate reflection of political events, which is critical for ensuring the credibility of the dataset (further details can be found in Section 3.2.1).
- Secondly, as explained in Section 3.2.2, we leveraged public APIs provided by the National Institute of Statistics of Spain to generate a diverse set of fact-based news articles. These APIs gave access to a vast array of objective statistical data on a variety of politically relevant matters, such as economic performance, unemployment rates, healthcare metrics, and more. By using this wealth of data, we were able to produce news content that reflected real-world political and socio-economic trends, ensuring

that our dataset maintained accuracy and relevance in terms of the key political issues that shape public discourse.

- Finally, due to the lack of publicly available repositories of political fake news in Spanish, we employed a hybrid approach to generating misleading content. This approach involved two distinct methods. The first method utilized the generative capabilities of AI, particularly a Generative Pre-trained Transformer (GPT) model, to automatically create a multitude of varied headlines. By using the OpenAI library, we generated fake headlines that closely mimicked the style and structure of real news articles (Section 3.2.3). The second method involved making strategic alterations to real news articles to create misleading information. These alterations typically included changing the names of political leaders or parties, as well as modifying key figures such as unemployment rates or dates, thereby creating plausible but false representations of political events. More detailed explanations of these methods are provided in Section 3.2.4.

3.2.1. Web Scraping

A portion of the authentic news in our dataset is sourced from *El Independiente* (www.elindependiente.com, accessed on 1 November 2023) and *Público* (www.publico.es, accessed on 1 November 2023); both are online media outlets renowned for their reputation as trustworthy and independent sources. They aim to avoid political bias and strive to deliver balanced and accurate coverage across various topics, making them robust choices for those seeking reliable news within the Spanish media landscape.

From a technical standpoint, we employed several tools and libraries to efficiently gather and process the data from both media outlets. First, we used a Python script to automate the process of accessing information from the websites. This script handled the task of systematically retrieving the web pages that contained the news articles we needed. To make the HTTP requests and retrieve the raw HTML content of these web pages, we utilized the requests library [46] to establish a connection with the server and navigate the various headers or parameters required by the media outlets' websites. Next, Beautiful Soup [47] was employed to handle the parsing of the raw HTML returned by requests, providing the necessary tools to navigate through each page's structure, extracting fields related to the headline, description, and publication date, as shown in the brief code snippet of Figure 2. This combination of requests and Beautiful Soup provided a reliable and efficient method for gathering structured data from various web pages, transforming raw HTML into clean, usable datasets ready for further processing.

```
URL = URLshort + str(page)

soup = BeautifulSoup(page.content, "html.parser")
elementsTitle = soup.find_all("h2", class_="entry-title")
elementsDate = soup.find_all("time", class_="entry-date")
elementsDescription = soup.find_all("div", class_="entry-subtitle")
```

Figure 2. Example of the utilization of the Beautiful Soup library to extract headlines, publication dates, and descriptions of news articles through web scraping.

3.2.2. Utilization of APIs Provided by the National Institute of Statistics

In order to enhance our dataset with precise and up-to-date information on politically and economically relevant issues, we have leveraged the valuable tools provided by Spain's National Institute of Statistics. Among these tools, we have access to public APIs that allow us to tap into a wide range of statistically objective data with clear political relevance, such as the unemployment rate, the country's economic growth within a specific period, the gross domestic product, inflation rate, public debt, and the consumer price index, to name just a few examples.

The requests made to the INE API were once again parsed using the Beautiful Soup library. From the extracted responses in XML format, we could, for instance, obtain highly diverse politically relevant information such as data about the average salary in Galicia, a region located in the northwest of Spain. This resulted in program-generated headlines such as "In 2020, the average full-time salary in Galicia reached 2136.4 euros", which were incorporated into our dataset. (Please note that the sample headline has been translated into English for the readers' convenience, while the original version in our dataset is in Spanish.)

3.2.3. Leveraging Generative Models via the OpenAI API

As previously mentioned, the incorporation of both authentic and deceptive news into the training dataset, with closely aligned language and structure, enables us to enhance the model's capability to discern very subtle distinctions between genuine and misleading information. For that purpose, we have resorted to generative large language models to reformulate the headlines scraped from reliable media reports (Section 3.2.1) and constructed from objective statistics retrieved from the INE API (Section 3.2.2). To do so, we have made different requests to OpenAI API, starting from specific headlines (e.g., "The Spanish government increases social spending in the 2022 state budget") to generate the different types of news, which are detailed next:

- First, we created news items that are equivalent to the original ones but with very similar wording (in our particular example, headlines such as "The Spanish government boosts social expenditure in the 2022 state budget", "The Spanish government increases in allocations for social initiatives in the 2022 state budget", or "The Spanish government amplifies social spending in 2022").
- Second, we crafted alternative news with different wording (e.g., "Spanish fiscal plan for 2022 boosts social expenditure", "The government of Spain boosts investment in social programs in 2022", "Spain allocates additional resources to social policies in 2022", "Spain increases budget allocations for social programs in 2022", "Spanish 2022 fiscal plan sees a surge in allocations toward social welfare initiatives", or "Spain allocates substantial increase to social investment in the 2022 fiscal plan").
- Third, we generated contrasting news using different terms (e.g., "In 2022 Spain decreases investment in social programs", "The Spanish government invests less in welfare policies in 2022", or "Contraction in Spain's social expenditure for 2022").
- Lastly, we also considered the generation of contrasting news using very similar terms (e.g., "The Spanish government reduces social spending in the 2022 state budget" or "The Spanish government advocate for reduction social spending in the 2022 state budget").

As depicted in Figure 3, we resorted to a GPT-3 model to create a multitude of fictitious news articles about economic data, healthcare, employability, etc., with varying semantic nuances and syntactic structures to enrich the knowledge from which the Transformers models trained with our custom corpus will learn [48]. Specifically, we employed the *gpt-3.5-turbo* model, which operates within a token limit of 4097. This limit encompasses both the input (the prompt given to the model) and the output (the response generated), meaning that the total number of text units exchanged in each interaction must stay within this boundary. Our approach involved interacting with the model by requesting batches of news generation until the token limit was reached, after which the content was categorized and saved appropriately. As shown in Figure 3, the roles in this process are defined as 'user' for providing prompts and requesting news, and 'assistant' for the model's response generation.

```
content = "Provide me with 100 fictional headlines in Spanish covering the topics of
economy, healthcare, education, and politics in Spain for the year 2022."

messages.append({"role":"user", "content": content})

completion = openai.ChatCompletion.create(
    model = "gtp-3.5.turbo",
    messages = messages
)

chat_response = completion.choices[0].message.content
writeFile.write(chat_response)
messages.append({"role":"assistant","content":chat_response})
```

Figure 3. Using the OpenAI API to craft fictional political news articles in the Spanish language.

3.2.4. Altering Real News to Generate Fake News

With the aim of further expanding the set of fake news within our synthetic corpus and, simultaneously, balancing both categories, we have chosen to modify certain data in the genuine news collected through the methods described in the preceding sections. Our intention is to change the names of political parties, political leaders, locations, and even numerical values in real news to make them appear fictitious, carefully supervising the resulting news items to ensure they are indeed fake.

- Locations such as cities, provinces, and countries, along with the names of political leaders, were identified using the spaCy library in conjunction with a publicly available pre-trained model specifically designed for named entity recognition (NER) tasks [49]. This powerful NER model enabled the accurate identification and classification of named entities within text, focusing on detecting and categorizing places and individuals. In our case, this approach played a crucial role in distinguishing regions, cities, and political figures in real news articles, allowing us to substitute these entities with names of the same category to generate corresponding fake news. Some examples are illustrated in Table 2. Given that the names used in this process were of Spanish origin, we employed the *es_core_news_lg* [50] model, tailored specifically for this language. In this way, we bypassed the need for other models, such as BERT for NER tasks; having been trained in English, these did not perform adequately in identifying the names of cities and even politicians originating from specific regions of Spain, such as Catalonia or Galicia.

Table 2. Examples of fake news created by manipulating various types of data within genuine political news articles.

Altered Data	True News	False News
Places	"**Catalonia** prepares for regional elections amid a pandemic and political crisis (February 2021)."	"**Asturias** prepares for regional elections amid a pandemic and political crisis (February 2021)."
Political leaders	"**Pedro Sánchez** announces the extension of permanent ERTEs until May 2021 (January 2021)."	"**Josep Borrell** announces the extension of permanent ERTEs until May 2021 (January 2021)."
Political parties	"**Podemos** defended the right to housing against speculation (March 2020)."	"**Partido Popular** defended the right to housing against speculation (March 2020)."
Numeric values	"The unemployment rate in Spain drops to **15.3%** in the second quarter of 2021 (July 2021)."	"The unemployment rate in Spain drops to **5.3%** in the second quarter of 2021 (July 2021)."

- The names of political parties have been manually compiled in a list, located within the real news, and replaced with other options in the fake news generated from the originals, as shown in Table 2.
- Numerical values have been identified with the assistance of the *Re* library [51], which enables the detection of patterns within text. Once these patterns and the values to be altered have been located, they are replaced with other random values within a coherent range, depending on the type of data to be modified (percentage, year, etc.).

By following the four procedures explained from Sections 3.2.1–3.2.4, a total of 57,231 news articles were collected, resulting in a balanced corpus with 52% of true news and 48% of fake news. This dataset can be downloaded from https://www.kaggle.com/datasets/javieroterovizoso/spanish-political-fake-news (accessed on 1 October 2024).

4. Tailoring Pre-Trained Transformer Models for Detection in Spanish

Modern NLP models are pre-trained on extensive text data using significant computational resources and large corpora, enabling them to gain a deep understanding of language. These pre-trained models can be fine-tuned for specific natural language tasks, involving the adjustment of hyperparameters that play a crucial role in their performance. While model parameters are weight matrices that capture relationships between words, phrases, and documents in the training corpus, hyperparameters impact the model's behavior during both training and inference, significantly affecting accuracy and capabilities.

In Section 4.1, we outline the fundamental features of the pre-trained Transformers explored in our approach, focusing on BERT, due to its outstanding performance, along with its primary optimized variations. Next, Section 4.2 justifies the selection of publicly available models for our political fake news detection approach. Lastly, Section 4.3 focuses on the training and fine-tuning for the chosen models on the dataset described in Section 3, with a specific emphasis on the adjustment of model hyperparameters and the evaluation metrics adopted in our experiments.

4.1. An Overview of BERT and Its Optimized Variants

BERT (Bidirectional Encoder Representations from Transformers) is a language model architecture renowned for its proficiency in capturing long-range dependencies in sequential data. The foundation of BERT lies in its sophisticated self-attention mechanisms, enabling it to model word interactions within a sentence and to weigh the relevance of each word's context. Thanks to these properties, BERT generates contextualized word embeddings that depend on the entire surrounding context of each word in the sentence, thus capturing intricate language dependencies. To do so, BERT utilizes two primary pre-training strategies: the Masked Language Model (MLM) and Next Sentence Prediction (NSP). In the MLM process, BERT selectively masks words within sentences, prompting the model to predict hidden words by considering both their left and right contexts within the sentence. Additionally, NSP involves training BERT to discern whether pairs of sentences randomly selected from a corpus are contiguous or from distinct parts of the text, thereby helping the model capture meaningful relationships between sentences. The combination of these techniques enables BERT to comprehend the intricate nuances of language semantics and syntax.

In addition to its fine-tuning on various natural language understanding tasks, as highlighted throughout this section, BERT has been adapted to create multilingual models with support for a wide range of languages, including Spanish, which is the focus of our research on fake news detection.

The notable performance displayed by BERT in NLP tasks [52–54] has paved the way for a multitude of optimizations and variations, including AlBERT [55], DistilBERT [56], and RoBERTa [57], whose main differences lie in their optimization goals and methods:

- AlBERT, or A little BERT, aims to improve language models for increased efficiency and speed while making some trade-offs in performance. It focuses on resource efficiency and faster processing while still maintaining a performance level slightly below that of the original BERT model. This is particularly valuable for applications where speed and resource efficiency are crucial, such as mobile natural language processing [58,59]. AlBERT accomplishes this through the redistribution of model parameters and the integration of various optimization techniques, which are out of the scope of the paper.
- DistilBERT, also known as Distilled BERT, shares a similar aim of enhancing language models by prioritizing efficiency and speed. It achieves this goal by creating a more

compact and faster model while retaining much of the knowledge from the original BERT. DistilBERT is tailored for applications where lightweight and fast language models are essential, making it well suited for scenarios with limited computational resources [60,61].

- In contrast, optimizations in RoBERTa are focused on enhancing natural language processing performance. This is achieved by conducting more extensive pre-training, using a larger text corpus compared to BERT, encompassing a wide range of natural language tasks prior to fine-tuning for specific applications, and more effectively adjusting hyperparameters. These enhancements empower the model to capture more intricate and nuanced language patterns, ultimately leading to superior performance in a variety of natural language processing tasks [62,63].

While we acknowledge the importance of optimizing efficiency and resource management, our primary objective is to enhance the model's performance to achieve the highest precision in the NLP task of news article classification within the scope of politics and in the Spanish language, which have received little attention in the literature to date. Consequently, we opted to fine-tune both the BERT and RoBERTa models, with a particular emphasis on those with support for the Spanish language, as outlined in the next section.

4.2. Selecting Pre-Trained Models for Detecting Fake News in Spanish

In the Hugging Face repository, there are numerous BERT and RoBERTa models of different sizes and purposes. In this section, we justify the use of the ones we have chosen for our experiments on fake news detection in Spanish.

- Available BERT models. On one hand, there are two variants of BERT, known as *BERT-base* and *BERT-large*. Both are bidirectional models pre-trained on extensive text data in English, serving as the foundation for fine-tuning in a wide range of NLP applications. Specifically, *BERT-base* is the lighter version with fewer parameters (110 million), resulting in a smaller model size. In contrast, *BERT-large* is a more substantial and powerful iteration, boasting more parameters (340 million) and greater representational capacity. As a consequence, *BERT-large* proves more effective but also demands more computational resources, making it the preferred choice for complex natural language processing tasks. Moreover, in addition to the variations in size and capacity between *BERT-base* and *BERT-large*, both models come in *cased* and *uncased* versions. The selection between these variants should be influenced by the nature of the data and the specific requirements of each NLP task. *Cased* versions are better suited when preserving uppercase and lowercase distinctions is essential, such as in proper nouns and specialized contexts, for tasks like named entity recognition or machine translation [64–66]. Conversely, *uncased* versions convert all text to lowercase before training, which is beneficial when capitalization carries no semantic significance. These models find value in text classification, natural language processing in social media, and sentiment analysis in informal texts [67–69].
  Various language-specific versions of BERT models are also available. Two notable examples include *BERT-multilingual* and *BETO* models, both of which are offered in *base*/*large* and *cased*/*uncased* versions. On one hand, *BERT-multilingual* is a BERT variant pre-trained on a diverse set of languages, including Spanish, enabling it to capture general language representations across a wide range of linguistic backgrounds [70]. On the other hand, *BETO* distinguishes itself as a language-specific model explicitly trained for the Spanish language. It delivers outstanding performance across a diverse range of natural language processing tasks in Spanish, making it particularly well suited for applications like sentiment analysis, topic categorization, and various other forms of text classification [71,72].
- Available RoBERTa models. RoBERTa is available in *base* and *large* versions, with all models being *case* sensitive. Additionally, there are multilingual variants, such as the *xlm-roberta-base* model, which was designed as a cross-lingual language model based

on RoBERTa to operate effectively in 100 different languages and pre-trained on an extensive dataset of 2.5 TB of filtered CommonCrawl data [73].
Furthermore, there are specialized RoBERTa models that have been pre-trained on comprehensive Spanish-language corpora. For example, the *bertin-roberta-base-spanish* model that was developed as part of the BERTIN project [74] and utilizes the mC4 corpus [75], a multilingual variant of the C4 dataset supporting over 100 languages sourced from the public Common Crawl web scrape. This dataset includes a total of 6.6 billion pages, with 416 million of them in Spanish [76]. Additionally, the *roberta-base-bne* model was created in 2021 by the Barcelona Supercomputing Center on 570 GB Spanish-language high-quality documents extracted from the National Library of Spain (BNE) (The acronym *BNE* in the model name *RoBERTa-BNE* refers to the Biblioteca Nacional de España in Spanish (National Library of Spain in English). Its website is https://www.bne.es/en (accessed on 1 November 2023).) using the supercomputer MareNostrum4 [77].

In our approach to text classification in Spanish, our experiments involved both the base and cased versions of BERT and RoBERTa models (identified hereafter as *BERT-base* and *RoBERTa-base*, respectively). In addition to these, we also included their corresponding multilingual variants (denoted as *BERT-multilingual* and *RoBERTa-multilingual*, respectively), along with the models specifically pre-trained on Spanish corpora (*BETO*, *RoBERTa-BERTIN* and *RoBERTa-BNE*). This selection is guided by several critical factors:

- The inclusion of models trained in English serves as a benchmark, allowing us to quantify their inferior performance compared to the variants adapted specifically for Spanish. This comparative analysis provides valuable insights into the extent of the performance gap between the English-trained models and their Spanish-tailored counterparts.
- Base models strike an effective balance between computational resources and their capacity to capture information and represent language, aligning well with the resources available in our Google Colab environment.
- In light of the unavailability of *uncased* RoBERTa models, we have opted to use *cased* BERT models to ensure a fair comparison. The dataset format has been meticulously inspected to guarantee the correct usage of lowercase and uppercase letters.
- The dedicated support for the Spanish language ensures the versatility and efficiency required to attain effective text classification within this specific linguistic context.
- Last but not least, all the above models are primarily aimed at being fine-tuned on tasks such as sequence classification, making them useful for our fake news detection purposes.

4.3. Fine-Tuning: Adjustment of Model Hyperparameters and Evaluation Metrics

In order to fine-tune the previously mentioned BERT and RoBERTa models on our domain-specific dataset containing Spanish political news, we made use of the Transformer library [78], and employed cross-validation to address the potential issue of overfitting in the models we intended to fine-tune.

- In particular, cross-validation allows us to assess the performance of the learned models and ensure their robustness in generalizing to unseen data. It works by partitioning the dataset into multiple subsets or folds, where the model is evaluated on one subset after being trained on the other ones in each iteration, ultimately yielding an average performance metric. This process aids in detecting overfitting as it highlights any significant performance discrepancies across the folds. If the model exhibits varying performance on different subsets, it signals that it may be overfitting to the training data, as it struggles to generalize to new, unseen data. In our approach, we have set aside 20% of the data for testing the final model (*test data*), and on the remaining 80%, we have implemented cross-validation with 5 folds ($K = 5$). In each

fold or iteration, the model is trained on 80% of the available data (*training data*) and its performance is evaluated on the remaining 20% (*evaluation data*).

- Regarding the fine-tuning, this process also encompasses the refinement of hyperparameter values, whose appropriate configuration is critical to build effective models. Both BERT and its RoBERTa models share common hyperparameters due to their Transformer-based architecture and pre-training process: (i) the *learning rate* controls how much the model weights are updated after each training step; (ii) the *maximum sequence length* determines the maximum size of the input that the model can handle; (iii) the *batch size* defines the number of training examples processed in each iteration, affecting memory usage and training speed; (iv) the *number of epochs* indicates how many times the model goes through the entire training dataset; and (v) *regularization techniques*, such as L2 regularization or weight decay, that help prevent overfitting.
  To optimize the hyperparameters, we carried out a series of experiments, adjusting their values based on the unique characteristics of our domain-specific dataset and the specific requirements of our binary text classification task. In doing so, we managed to strike a balance between achieving high performance on the training data and ensuring effective generalization to unseen data for each model. We carefully tuned the hyperparameters of the Transformer models evaluated through an extensive series of experiments, exploring different combinations until identifying the settings that offered the best performance. This iterative process allowed us to determine the optimal values that maximize the models' accuracy. The chosen values for each type of model are presented in Table 3, where we also specify the seed used for reproducibility by consistently initializing random processes across model runs.
  Although tools such as GridSearchCV and RandomizedSearchCV, along with more sophisticated options like Optuna and Hyperopt, provide valuable techniques for hyperparameter tuning, they have notable drawbacks in our particular scope. GridSearchCV systematically examines a specified grid of parameters, but it can become resource intensive as the number of parameters increases. In contrast, RandomizedSearchCV enhances efficiency by randomly selecting parameter values but may overlook significant combinations. In light of these challenges, we chose to pursue a manual tuning approach, which offers the flexibility to concentrate on essential parameters based on our expertise in the field, thus avoiding unnecessary computational demands. Unlike automated techniques, manual tuning facilitates iterative adjustments of parameters informed by practical observations, allowing for a more targeted and effective investigation of the hyperparameter landscape. This method minimized the likelihood of missing crucial settings and granted greater control over the optimization process, ultimately resulting in outcomes that aligned more closely with our specific requirements.

Table 3. Hyperparameter values that have yielded the best performance in the 3 BERT models and 4 RoBERTa models fine-tuned in our approach.

Hyperparameters	BERT Models	RoBERTa Models
	BERT-Base ***BERT-Multilingual*** ***BETO***	***RoBERTa-Base*** ***RoBERTa-Multilingual*** ***RoBERTa-BERTIN*** ***RoBERTa-BNE***
Learning rate	5×10^{-5}	10^{-5}
Max sequence length	512	128
Batch size	10	16
Epochs	10	10
Drop out rate	0.1	0.15
L2 regularization	0.005	0.001
Seed	10	16

- Lastly, during the fine-tuning process, it is essential to measure various metrics to evaluate the performance of each model trained on our Spanish political fake news dataset. In Table 4, we describe the metrics adopted in our approach, obtained using the *Sklearn* library [79], where *True Positives* (TP) represent instances that are correctly predicted as positive (real news) by the model, while *False Positives* (FP) are instances that the model incorrectly predicts as genuine news when they are, in fact, false. Similarly, *False Negatives* (FN) occur when the model incorrectly predicts instances as negative (fake news) that are truly positive (real news), and *True Negatives* (TN) are instances that are correctly predicted as deceptive news by the model.

Table 4. Well-known evaluation metrics adopted in our approach.

Metric	Definition	Mathematical Equation
Precision	Proportion of genuine news correctly predicted by the model.	$\frac{TP}{TP+FP}$
Recall	Ratio of all existing authentic news articles correctly classified by the model.	$\frac{TP}{TP+FN}$
Accuracy	Proportion of right predictions, regardless of whether they were positive or negative, made by the model.	$\frac{TP+TN}{TP+TN+FP+FN}$
F1-score	Evaluates how well the model balances precision and recall, that is, it focuses on how many of the predicted positives are TP and how many of the TP are captured.	$2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$
MCC	Evaluates the performance considering TP, TN, FP and FN to provide a comprehensive assessment of model prediction quality (1: perfect predictions, 0: random predictions, and -1 entirely inaccurate predictions).	$\frac{TN \cdot TP - FN \cdot FP}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$

Although AUC-ROC (Area Under the Curve) is a widely used metric in binary classification, high values may not be representative in our scenario, where we work with synthetic and non-synthetic data. In this context, elevated AUC-ROC values would suggest strong performance in distinguishing real from fake news, but this could be unrealistic due to the possibility that the synthetic nature of the fake news might make these samples easier for the model to detect, potentially leading to results that do not generalize well to real-world settings.

While AUC-ROC is often useful for balanced datasets, it can overlook specific types of errors, giving an overly optimistic impression of the model's effectiveness. For these reasons, we prioritized metrics that provide a more detailed evaluation of error management. Specifically, by focusing on precision and recall, we evaluate the model's ability to balance the trade-off between false positives (misclassifying real news as fake) and false negatives (failing to identify fake news), ensuring robust error handling. The F1 score further strengthens this assessment by offering a balanced metric that combines both precision and recall, while MCC offers a broader view of the model's performance by accounting for all types of outcomes.

Before focusing on the results and experimental discussion, we emphasize the fine-tuning of the evaluated models, detailing the process where key concepts discussed in this section, such as hyperparameter tuning and cross-validation, are integrated. The fine-tuning and testing codes are available on GitHub at https://github.com/gssi-uvigo/debatrue (accessed on 1 October 2024).

- Initially, the pre-trained models are loaded using the relevant Hugging Face classes (e.g., RobertaForSequenceClassification for the RoBERTa-BNE model), and the number of labels is set to two for our fake news detection task. The entire model is fine-tuned without freezing any layers, allowing all layers to be optimized during training to fully adjust to the binary classification task.
- To ensure that the model generalizes well to unseen data, we use 5-fold cross-validation. Therefore, the dataset is divided into five subsets, training the model on four subsets and validating on the remaining one. During each training phase of the folds,

we use the AdamW optimizer with a learning rate of 10^{-5} and L2 regularization to enhance weight decay and mitigate overfitting. After each of the ten epochs, we compute the cross-entropy loss between the predicted outputs and the true labels to evaluate the model's performance. Gradients are computed using backpropagation (*loss.backward()*), and the model's parameters are updated (through *optimizer.step()*). This process is repeated for each batch within the epoch, gradually refining the model's parameters.

- Lastly, metrics such as precision, recall, accuracy, F1 score, and MCC are computed after each epoch to measure the model's performance, ensuring a reliable evaluation.

5. Results and Experimental Discussion

In this section, we present the results of our experimental evaluation conducted on the seven models introduced in the previous sections. As depicted in Table 3, these models include the two base models of BERT and RoBERTa, pre-trained in English (denoted as *BERT-base* and *RoBERTa-base*, respectively), as well as their multilingual versions (*BERT-multilingual* and *RoBERTa-multilingual*). Additionally, we consider the three models specifically pre-trained on extensive collections of texts in Spanish: *BETO*, *RoBERTa-BERTIN*, and *RoBERTa-BNE*. Notice that we have reserved 20% of the data from our dataset of 57,231 news articles for testing, employing a 5-fold cross-validation approach on the remaining 80%.

Given the significant advantages demonstrated by DL models, particularly Transformer-based models, in capturing complex patterns and long-range dependencies within data, we chose to focus on DL methods rather than revisiting traditional ML approaches [18–20]. As highlighted in Section 2, ML models depend on manual feature engineering, which can miss complex patterns and deeper context, and often assume linear relationships, limiting their ability to capture intricate data interactions. In contrast, DL models automatically learn both simple and complex features from raw data, making them more effective for large and varied datasets. In this regard, studies conducted in various languages, such as English [23], Greek [44], and Bengali [45], consistently demonstrate the effectiveness of BERT models and their variants for fake news detection. These studies underline the ability of Transformer-based models to capture subtle language patterns and rich contextual information, leading to superior performance over traditional models in text classification tasks. By utilizing self-attention mechanisms, these models are able to understand complex relationships between words, making them particularly suited for identifying misinformation. Given these findings, we have chosen to focus our experiments on fine-tuned BERT and RoBERTa models, as these Transformer architectures have shown to be highly effective in addressing the challenges specific to this domain.

We present our results as follows: firstly, we compare the performance of fine-tuned BERT and RoBERTa models during training (Section 5.1) and on unseen data (Section 5.2). Additionally, beyond testing on 20% synthetic Spanish political fake news, we experimented with a smaller, independent dataset of manually curated real and non-synthetic fake news (Section 5.2.1). This aimed to assess the models' ability to detect authentic political fake news in Spanish, expanding beyond the synthetic dataset used for fine-tuning. Finally, in Section 5.2.2, we benchmark our approach against related methods in general text classification and specific fake news detection across multiple languages.

5.1. Assessing Loss and Accuracy Metrics during Model Fine-Tuning

Our experimental evaluation tracks the evolution of training losses and the evaluation accuracy of each model across multiple epochs, enabling us to identify irregularities that may indicate a performance gap between the training and evaluation sets during cross-validation, potentially signaling overfitting.

Figure 4 provides evidence that the training losses of the three fine-tuned BERT models decrease as the number of training epochs increases and stabilizes after reaching convergence. This observation indicates their effective adaptation to the domain-specific training data within our methodology. Notably, the BETO model converges more rapidly

than the others, which aligns with expectations, given its extensive pre-training on a large Spanish-language corpus and subsequent fine-tuning on our dataset in the same language.

Train Loss										
Fine-tuned model	1	2	3	4	5	6	7	8	9	10
BERT-base	0.2383	0.1122	0.0759	0.0487	0.0337	0.0240	0.0190	0.0172	0.0165	0.0167
BERT-multilingual	0.1733	0.0771	0.0436	0.0268	0.0183	0.0141	0.0116	0.0099	0.0081	0.0071
BETO	0.1723	0.0689	0.0345	0.0191	0.0132	0.0103	0.0084	0.0068	0.0065	0.0059

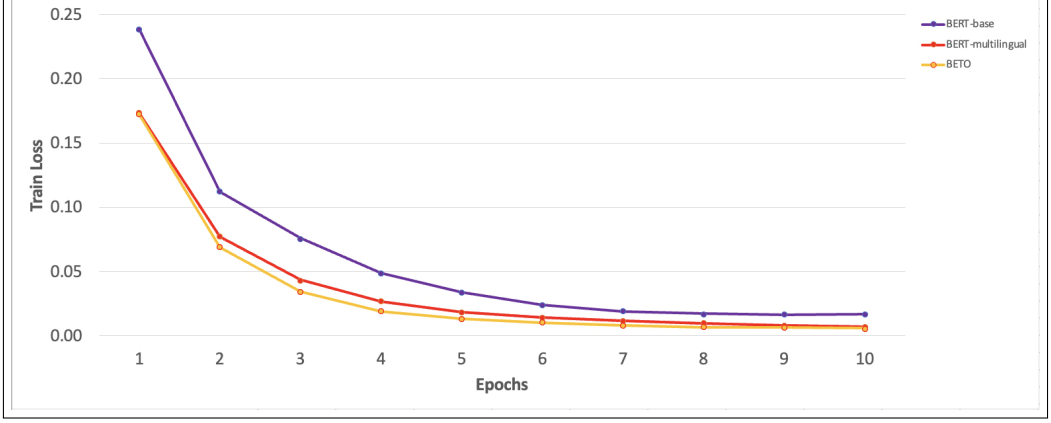


Figure 4. Progression of training loss over training epochs for the 3 fine-tuned BERT-based models using 80% of our Spanish political fake news dataset.

An analysis of the trends in evaluation accuracy (Figure 5) over multiple epochs confirms that the BETO model offers the best overall performance. This model consistently achieves the highest and most reliable accuracy. On the other hand, the BERT-multilingual model displays lower losses than the BERT-base model, which records the lowest accuracy, thus confirming its inferior performance in a classification task in a language (Spanish) different from its pre-training language (English). Since the evaluation accuracy values are high and stable as the BERT models train, it is evident that none of them are overfitting the training data, and they generalize well to unseen data during cross-validation.

Eval Accuracy										
Fine-tuned model	1	2	3	4	5	6	7	8	9	10
BERT-base	0.9559	0.9615	0.9667	0.9646	0.9587	0.9633	0.9639	0.9628	0.9587	0.9633
BERT-multilingual	0.9630	0.9713	0.9746	0.9709	0.9726	0.9733	0.9739	0.9722	0.9696	0.9699
BETO	0.9561	0.9752	0.9783	0.9733	0.9772	0.9793	0.9760	0.9765	0.9730	0.9726

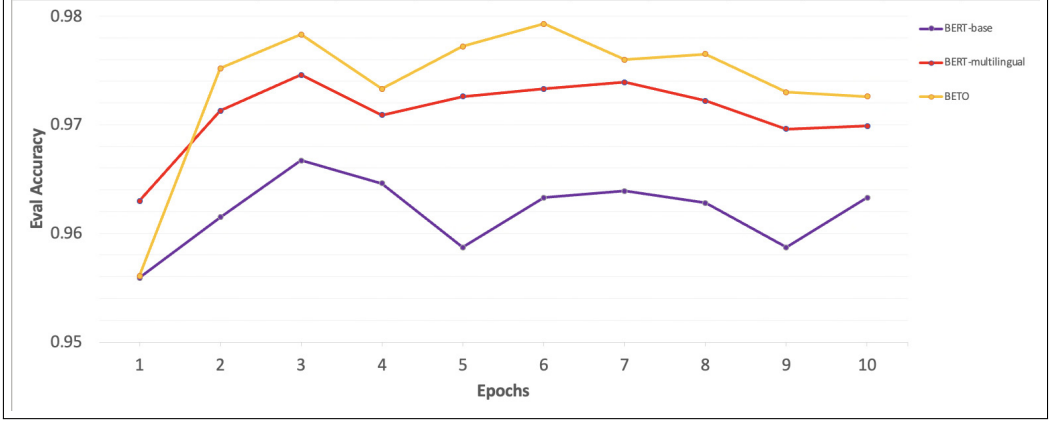


Figure 5. Progression of evaluation accuracy over training epochs for the 3 fine-tuned BERT-based models using 80% of our Spanish political fake news dataset and cross validation.

The positive influence of pre-training with Spanish text corpora is also evident when examining the training and evaluation metrics of RoBERTa-based models. In particular, both training losses (Figure 6) and evaluation accuracy (Figure 7) consistently stabilize, which confirms the absence of overfitting.

Train Loss										
Fine-tuned model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
RoBERTa-base	0.2690	0.1130	0.0732	0.0476	0.0338	0.0281	0.0224	0.0184	0.0198	0.0151
RoBERTa-multilingual	0.2319	0.0863	0.0543	0.0351	0.0230	0.0193	0.0150	0.0130	0.0114	0.0105
RoBERTa-BNE	0.1145	0.0299	0.0157	0.0106	0.0077	0.0054	0.0062	0.0048	0.0036	0.0044
RoBERTa-BERTIN	0.1645	0.0471	0.0233	0.0140	0.0108	0.0100	0.0033	0.0032	0.0040	0.0003

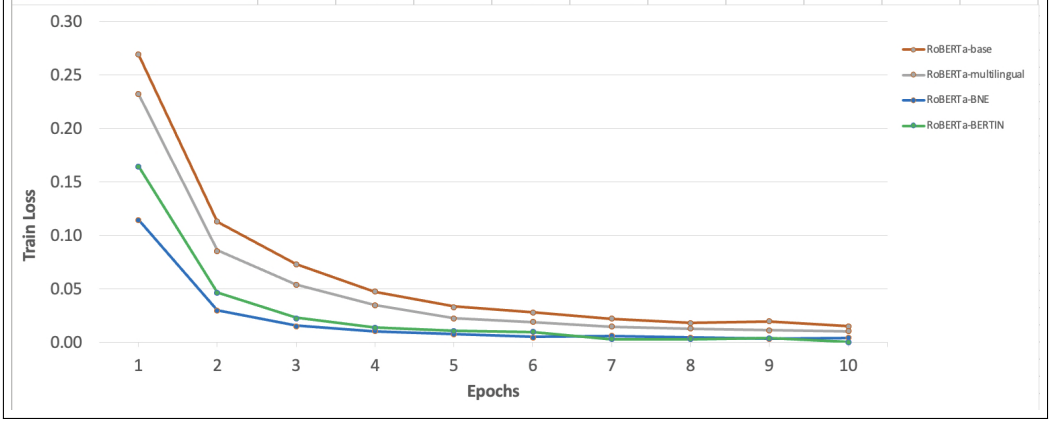


Figure 6. Progression of training loss over training epochs for the 4 fine-tuned RoBERTa-based models using 80% of our Spanish political fake news dataset and cross validation.

Eval Accuracy										
Fine-tuned model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
RoBERTa-base	0.9533	0.9652	0.9678	0.9657	0.9661	0.9620	0.9674	0.9670	0.9683	0.9635
RoBERTa-multilingual	0.9696	0.9720	0.9737	0.9776	0.9728	0.9763	0.9765	0.9735	0.9750	0.9757
RoBERTa-BNE	0.9730	0.9835	0.9822	0.9843	0.9833	0.9815	0.9813	0.9833	0.9846	0.9830
RoBERTa-BERTIN	0.9700	0.9800	0.9762	0.9780	0.9791	0.9790	0.9789	0.9778	0.9800	0.9811

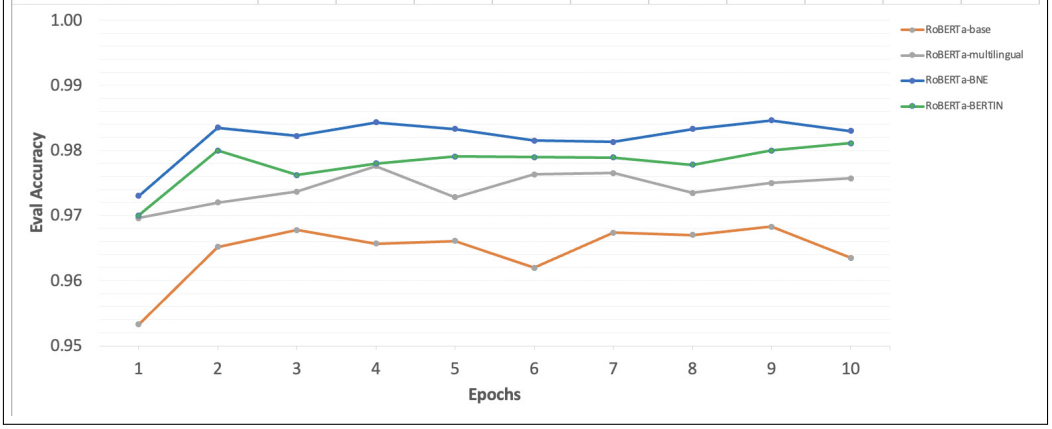


Figure 7. Progression of evaluation accuracy over training epochs for the 4 fine-tuned RoBERTa-based models using 80% of our Spanish political fake news dataset and cross validation.

The *RoBERTa-BNE* model consistently achieves the highest evaluation accuracy, with *RoBERTa-BERTIN* following closely, as depicted in Figure 7. However, the *RoBERTa-multilingual* model, despite its inclusion of Spanish among the supported languages, does not attain the performance levels of the first two models. Nevertheless, it exhibits a noteworthy enhancement when contrasted with the base model, which was originally pre-trained on English text corpora.

To provide a more transparent comparison among the five fine-tuned models from BERT and RoBERTa families that have performed better (multilingual and Spanish-specific models), their training and evaluation metrics are collectively presented in Figures 8 and 9.

Train Loss										
Fine-tuned model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
RoBERTa-multilingual	0.2319	0.0863	0.0543	0.0351	0.0230	0.0193	0.0150	0.0130	0.0114	0.0105
RoBERTa-BNE	0.1145	0.0299	0.0157	0.0106	0.0077	0.0054	0.0062	0.0048	0.0036	0.0044
BERT-multilingual	0.1733	0.0771	0.0436	0.0268	0.0183	0.0141	0.0116	0.0099	0.0081	0.0071
BETO	0.1723	0.0689	0.0345	0.0191	0.0132	0.0103	0.0084	0.0068	0.0065	0.0059
RoBERTa-BERTIN	0.1645	0.0471	0.0233	0.0140	0.0108	0.0100	0.0033	0.0032	0.0040	0.0003

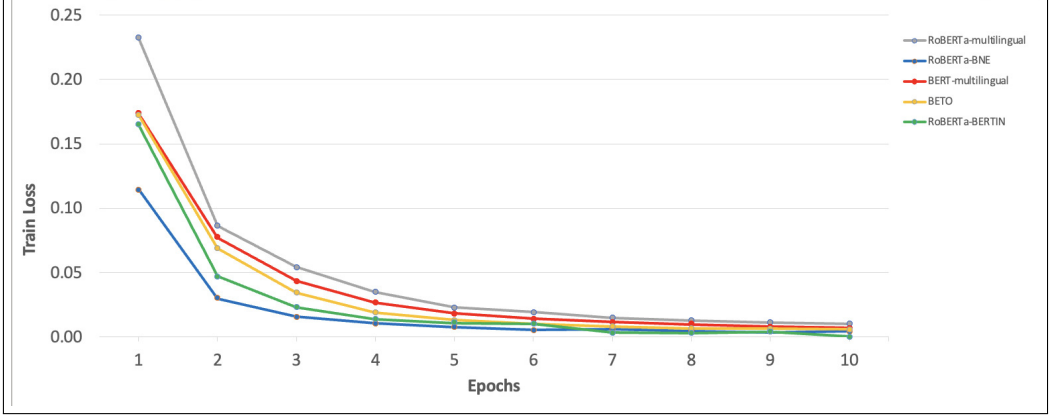


Figure 8. Comparing the progression of training losses across epochs for fine-tuned BERT-based and RoBERTa-based models.

Eval Accuracy										
Fine-tuned model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
RoBERTa-multilingual	0.9696	0.9720	0.9737	0.9776	0.9728	0.9763	0.9765	0.9735	0.9750	0.9757
RoBERTa-BNE	0.9730	0.9835	0.9822	0.9843	0.9833	0.9815	0.9813	0.9833	0.9846	0.9830
BERT-multilingual	0.9630	0.9713	0.9746	0.9709	0.9726	0.9733	0.9739	0.9722	0.9726	0.9728
BETO	0.9561	0.9752	0.9783	0.9733	0.9772	0.9793	0.9720	0.9765	0.9730	0.9726
RoBERTa-BERTIN	0.9700	0.9800	0.9762	0.9780	0.9791	0.9790	0.9789	0.9778	0.9800	0.9811

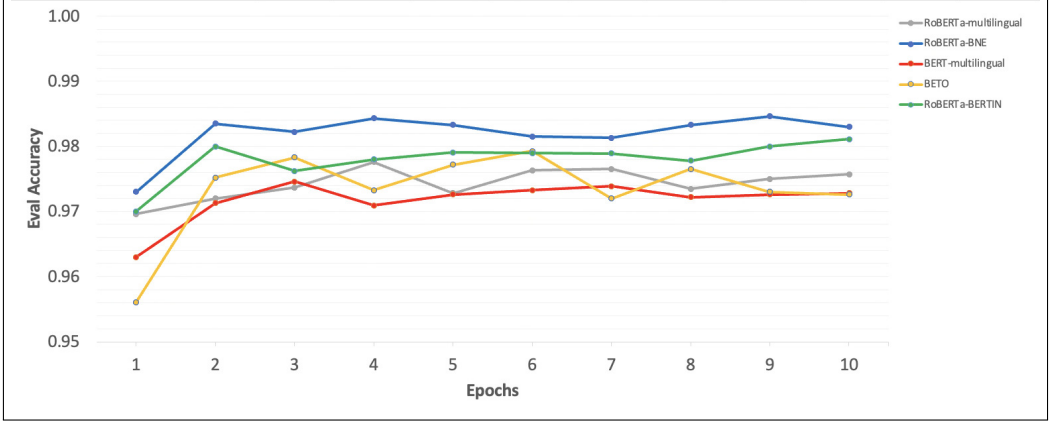


Figure 9. Comparing the progression of evaluation accuracy across epochs for fine-tuned BERT-based and RoBERTa-based models.

The data illustrated in both graphs unveil the superiority of the RoBERTa models over the BERT variants, as anticipated due to the enhancements introduced in RoBERTa building upon the original BERT architecture. Notably, the Spanish adaptations of RoBERTa, such as BNE and BERTIN, outperform BETO, the Spanish BERT. Furthermore, in line with our earlier discussion, *RoBERTa-BNE* demonstrates the best performance during the fine-tuning process on our dataset, tailored specifically for Spanish political fake news detection.

After assessing the behavior of the BERT and RoBERTa model groups throughout this section during the learning phase, our focus shifts to evaluating their respective metrics on previously unseen data, which constitute 20% of our dataset and were not utilized in the training process. This evaluation enables us to gauge the models' generalization capabilities and measure their performance on novel instances, providing valuable insights into their effectiveness beyond the training dataset.

5.2. Analyzing Test Metrics of the Fine-Tuned Models

The models analyzed in our study are evaluated using five well-known metrics: accuracy, precision, F1 score, recall, and MCC. We present the results achieved by the models without fine-tuning in Table 5, and those corresponding to their fine-tuned versions on our domain-specific dataset in Table 6.

Table 5. Metrics assessed during testing of the base BERT and RoBERTa models, without fine-tuning, using previously unseen data, which constitute 20% of our Spanish political fake news dataset.

Fine-Tuned Model	Test Accuracy	Test Precision	Test F-1 Score	Test Recall	Test MCC
BERT-multilingual	0.5942	0.5944	0.7452	0.9985	0.0057
BETO	0.4855	0.6881	0.3670	0.2456	0.0991
RoBERTa-multilingual	0.5942	0.5942	0.7455	1	0
RoBERTa-BERTIN	0.5489	0.6368	0.5963	0.5607	0.0908
RoBERTa-BNE	0.4383	0.5938	0.2687	0.1737	0.0005

The results depicted in Table 5 show that the pre-trained models, without fine-tuning, exhibit poor performance, with metrics suggesting behavior close to random, as indicated by MCC values near zero. This can be explained by the fact that base models are generally trained on extensive but generalized corpora. While this gives them a broad understanding of language, they lack the specialized knowledge required for our binary classification task. In our case, the dataset provides domain-specific information, particularly in distinguishing between fake and real news in the context of national politics. Without fine-tuning, the models are unable to grasp the nuanced patterns needed for accurate classification, as shown by the experimental results.

On the contrary, the values of Table 6 corroborate the notable capabilities of all fine-tuned variants, with indicators exceeding 0.9, and also reaffirm the superior performance of *RoBERTa-BNE* compared to its counterparts. In the context of fake news detection, the model's accuracy score of 0.986 signifies its ability to accurately classify over 98% of news articles as either genuine or fake. Precision, F1 and recall scores, ranging between 0.985 and 0.991 highlight the model's proficiency in not only reducing false positives but also effectively identifying a significant proportion of actual fake news instances. Furthermore, the MCC value of 0.971 underscores the model's robustness, as it can make precise predictions while taking into account both false positives and false negatives. This establishes the *RoBERTa-BNE* model as an certainly reliable choice for fake news detection, particularly in the domain of politics and in Spanish, two areas that have been relatively underexplored in the state-of-the-art research. The metrics also reveal that the *RoBERTa-BERTIN* model outperforms *BETO*, although the latter exhibits slightly better performance on previously unseen data compared to the multilingual versions of BERT and RoBERTa.

Table 6. Metrics assessed during testing of the fine-tuned BERT-based and RoBERTa-based models using previously unseen data, which constitute 20% of our Spanish political fake news dataset.

Fine-Tuned Model	Test Accuracy	Test Precision	Test F-1 Score	Test Recall	Test MCC
BERT-multilingual	0.925	0.922	0.930	0.938	0.911
BETO	0.927	0.927	0.931	0.935	0.913
RoBERTa-multilingual	0.924	0.921	0.928	0.935	0.912
RoBERTa-BERTIN	0.931	0.930	0.934	0.940	0.928
RoBERTa-BNE	**0.986**	**0.985**	**0.988**	**0.991**	**0.971**

The results obtained in the previous sections confirm that our fine-tuned Transformer-based models achieve a very notable performance in identifying fake news in the dataset of Spanish political news we have collected. However, since this dataset combines natural true news items with synthetically generated fake ones, it is possible that the fine-tuned models have learned to distinguish between the two types of data rather than learning to detect fake news, which is the goal of our research. This distinction might arise from subtle

differences in token distributions, which a linguistic model like BERT and its optimized version RoBERTa could pick up on. Although our experimental procedure was correctly applied, with a separate test split of 20%, this split was derived from the same underlying generating process. Therefore, there might be no assurance that the results will scale to actual fake news detection.

To address this issue, we conducted an additional evaluation using an independent dataset that is described in the next section.

5.2.1. Testing with a Dataset of Real and Fake News Naturally Sourced from Web Sources

The new dataset comprises 63 headlines extracted from various websites in July 2023. These headlines were collected from online media outlets that curated statements made by leaders of the two main political parties following a significant electoral debate during the Spanish general elections that month. These media outlets identified the statements as either true or false, providing contextual explanations to justify their classification. This process allowed us to create a non-synthetic dataset containing both real and fake news. A sample of the online media publications used in this process can be found at https://elpais.com/espana/elecciones-generales/2023-07-10/mentiras-y-medias-verdades-del-debate-de-atresmedia-entre-sanchez-y-feijoo.html?ssm=whatsapp (accessed on 1 November 2023). Out of the 63 headlines, 33 were labeled as fake, while 30 were classified as genuine. All the headlines were manually reviewed to ensure that the articles were correctly categorized as either fake or real news.

The evaluation was conducted using the RoBERTa-BNE model, which achieved the best performance according to the metrics outlined in Table 6. We tasked this model with classifying the 63 headlines from the new dataset as either fake or real news, and it achieved an accuracy rate of 71%. This represents a notable decline compared to the accuracy of over 98% observed when testing the model on the original synthetic dataset.

This drop in accuracy can be explained by several factors related to the differences between the datasets. The initial synthetic dataset, on which the model was trained, generated fake news by altering names, locations, and figures, or using tools like ChatGPT. The synthetic dataset also included true news articles sourced from online newspapers and INE, covering political issues such as education, the economy, employment, healthcare, and culture. However, the real news articles in the synthetic dataset did not comprehensively cover much of the information found in the non-synthetic dataset. The latter primarily consisted of statements made by political leaders during electoral debates leading up to the Spanish general elections in July 2023, as reported by various media outlets, from which we extracted the headlines.

In view of the obtained results, the synthetic nature and specific focus of the initial dataset might have provided the evaluated Transformer models with a consistent and predictable pattern to learn from. In contrast, the new (non-synthetic) dataset consisted of small authentic news articles from a variety of web sources, presenting a broader and more realistic range of fake news scenarios. The diverse and less predictable nature of real-world data posed additional challenges for the Transformer-based model, revealing its limitations in handling the complexities and variations inherent in genuine news articles. Furthermore, there were instances of mixed news in the dataset where candidates' statements combined both factual claims and false information within the same declaration. The presence of such mixed news likely introduced an extra challenge for the models, as they had to distinguish between true and false elements within a single piece of news. This complexity probably contributed to the observed decrease in accuracy, underscoring the model's limitations in dealing with more nuanced, real-world data.

To sum up, our evaluation shows that while the model demonstrates significant potential, its application in real-world scenarios requires further refinement to effectively manage the complexities and nuances of authentic fake news. Despite this limitation in our current approach, the results are encouraging, particularly in the relatively unexplored area of detecting political fake news in Spanish.

5.2.2. Performance Benchmark with Related Approaches

To conclude our experimental validation, we present a comparison between our leading model, *RoBERTa-BNE*, and various models discussed in Section 2. In this benchmarking, we have considered the two following criteria when selecting recent relevant approaches: those employing Transformers, whether for fake news detection specifically or for general text classification across multiple languages; and works that specifically focus on fake news detection in Spanish, incorporating both Transformer-based and traditional ML approaches.

It is important to note that our models and those referenced in the literature for benchmarking were not evaluated on the same datasets for two primary reasons. First, although most datasets used in the literature are accessible through popular repositories, the corresponding model codes are often unpublished, preventing direct comparison with our fine-tuned Transformers. Second, a fair comparison would require evaluating our fake news detection approach in Spanish, focused on politics, against similar approaches in the same language and thematic area. However, most published works concentrate on generic news in English and do not specifically address the political topic. Therefore, in line with our previous comments, we have compiled results from relevant studies utilizing Transformers to provide a comprehensive view of the state of the art in fake news detection and to highlight our contributions to a relatively unexplored area: detecting political fake news in Spanish. The benchmarking results can be seen in Table 7.

Table 7. Comparing our approach to detecting fake news about politics in Spanish using Transformers with other relevant research on text classification across several languages.

Approach	Adopted Technique	Dataset	Size	Language	Accuracy
[12]	Random Forest XGBoost Logistic Regression	Fakes1000 [34]	2000	Spanish	74.25% 64.25% 77%
[14]	SVM Naive Bayes KNN (K = 3) Decision Tree	The Spanish Fake News corpus [35]	971	Mexican Spanish	81% 59% 81% 65%
[13]	Random Forest BERT	Unpublished dataset	634	Spanish	89.80% 95.80%
[33]	Two parallel BERT networks	Fakenewsnet [41]	23,000	English	85.4%
[42]	BETO	Corpus extracted from [34,43]	2500	Spanish	80%
[15]	BERT	EXIST [36]	13,000	English Spanish	76.83%
[16]	BERT	IMDB movie reviews [37] RealOrNot tweets [38] Portuguese news [39] Chinese hotel reviews [40]	50,000 10,000 12,145 6000	English English Portuguese Chinese	93.87% 83.64% 91.20% 93.81%
Our approach	A fine-tuned RoBERTa-BNE model.	The model has been fine-tuned and tested on our Spanish Politician Fake News synthetic dataset.	57,231	Spanish	**98.6%**
Our approach	The same fine-tuned RoBERTa-BNE model as in the previous row.	The model was fine-tuned on the synthetic dataset and later tested on a new non-synthetic fake news dataset, including political news from Spain's July 2023 electoral debates.	63	Spanish	**71%**

In essence, our model fine-tuned on the synthetic dataset has delivered highly promising results, showcasing a remarkable accuracy rate of 98.6%. Moreover, the other fine-tuned models on our dataset have also produced outstanding outcomes, with the lowest accuracy reaching 92.5%. Significantly, this metric not only evidences a considerable performance enhancement compared to approaches employing various ML techniques (rows 1 and 2 in Table 7) but also stands out in contrast to works utilizing Transformer models. Our

results also confirm that selecting a single fine-tuned RoBERTa-BNE model with the appropriate adjustments can lead to better outcomes compared to more intricate approaches, as illustrated in rows 3 to 5 of Table 7. These approaches involve the use of individual BERT models pre-trained on English corpora, fine-tuning models on Spanish data corpora (e.g., BETO), and even the combination of multiple BERT models tailored for the specific task of fake news detection, respectively.

Furthermore, our approach attains the top performance with regards to other existing approaches not only in the specific domain of fake news detection in Spanish (and even in other languages like English, Portuguese, or Chinese) but also in the broader field of automated text classification. This extends beyond fake news detection to encompass various areas, including hotel and movie review classification or the identification of sexist language in Twitter posts, among others, as indicated in the works listed in the last rows of Table 7. These outstanding results can be attributed to the substantial size, diversity, and representativeness of our dataset, enhancing the limited corpora available for fake news detection published to date.

As previously mentioned, the non-synthetic dataset used in our experiments, including only 30 authentic news articles from diverse web sources, introduces broader and more realistic fake news scenarios. The unpredictability of these real-world data highlights challenges in the model's capability to manage the inherent complexities of genuine news articles, as evidenced by a notable decrease in accuracy, to 71%. While this accuracy level may be considered acceptable, particularly in less-explored domains like detecting political news in minority languages such as Spanish, it underscores the necessity for further refinement to effectively address these complexities in practical applications.

6. Conclusions and Further Work

The approach outlined in this paper underscores the significance of Transformer models in identifying political fake news, particularly within the Spanish language, a domain that has received limited attention in current research. We contribute by creating a high-quality and well-balanced dataset, bridging a gap in publicly available datasets for Spanish fake news detection. This collection is regarded as a valuable resource for future research, simplifying the construction and evaluation of models in this specific domain, and addressing the challenge of countering misinformation in a widely spoken language that is currently underrepresented in public datasets.

The superior performance of our fine-tuned model compared to the original models can be attributed to several key improvements, particularly those associated with the RoBERTa and BERT models. RoBERTa introduces several optimizations over BERT, mainly focusing on training efficiency rather than changing the underlying architecture. These include training on a significantly larger dataset and for more steps. RoBERTa replaces BERT's static masking with dynamic masking, where the masked words change in each training iteration. It also removes the Next Sentence Prediction (NSP) task, which was found to add little value. Additionally, RoBERTa uses larger batch sizes, trains with longer sequences from the beginning, and fine-tunes hyperparameters more extensively. These improvements allow the model to be exposed to more data and enhance the adjustment of weights, ultimately enabling RoBERTa to capture the complexities of natural language more effectively, leading to better performance in tasks like fake news detection, as our results revealed.

Our approach also benefits from using multilingual models BERT and RoBERTa, as well as Spanish-specific models such as BETO, RoBERTa-BERTIN, and RoBERTa-BNE, which are crucial for handling the nuances of the Spanish language, an especially important aspect for detecting fake news in Spanish political contexts. Additionally, fine-tuning the models on political news from Spain provides them with domain-specific knowledge, enabling better differentiation between true and false information in this specialized area.

Even though generating synthetic fake news has potential risks, several important considerations have guided our decision to incorporate this approach into our research. Although there are challenges, such as token distribution differences between real and

synthetic data, synthetic data offer significant advantages in contexts like Spanish political news, where direct sources for extracting fake news are limited. The scarcity of datasets specifically for Spanish-language fake news led us to explore alternative methods for generating diverse training data. To ensure the credibility of our data, we sourced authentic news articles from reputable, independent outlets known for their impartiality and balanced reporting, which is crucial in the political domain. Regarding linguistic and stylistic consistency, our dataset is composed of formal news articles from established journalistic sources in Spain, where differences in language use across regions are minimal. This formal consistency between real and synthetic data minimizes concerns about regional variations affecting the model's performance. For the generation of synthetic fake news, we addressed potential risks by modifying only specific factual details, such as names, dates, or figures, while maintaining the original structure, style, and syntax of the real news. This ensures that the tokens remain largely consistent with real data, making it difficult for the model to distinguish between natural and synthetic news based on superficial linguistic patterns, thus enabling the model to focus on detecting misinformation patterns aligned with the goals of our study.

In addition to customizing models for a specific topic and language, we perform a comparative analysis to assess the effectiveness of our fake news detection approach compared to other relevant works. Our experiments confirm that the *RoBERTa-BNE* model achieves exceptional precision, exceeding 98% when fine-tuned on our dataset. This surpasses previous methods that relied on combining multiple BERT models with lower-quality datasets across different languages, including Spanish. However, testing with a much smaller independent dataset of manually curated real and non-synthetic fake news reveals a significant drop in performance to 71%. This decline underscores the need for further adjustments to effectively address real-world challenges in accurately detecting variations within authentic fake news data. Nonetheless, this result seems promising as an initial exploration into a less-explored research domain such as political news detection in Spanish.

In addition to these refinements, our future research will focus on several key areas.

First, we plan to explore ways to leverage user feedback to enhance the model's accuracy and usability. To gather this feedback effectively, we aim to implement a system that encourages users to share their experiences. One approach is to create user surveys that capture insights about their interactions with the fake news detection model, particularly regarding prediction accuracy and overall usability. Additionally, introducing a 'report error' button will enable users to directly flag misclassifications, simplifying the feedback process. By analyzing this input, we can identify patterns in misclassifications and make targeted improvements.

Second, we intend to investigate a multimodal approach to enhance political fake news detection in Spanish by integrating text and images, especially in social media contexts. By analyzing visual elements, such as infographics and misleading images, alongside textual narratives, we can gain deeper insights into misinformation dissemination. This integration will reveal inconsistencies that text analysis alone may miss. Leveraging convolutional neural networks (CNNs) for image analysis, combined with Transformer-based models discussed in this paper, can help identify patterns indicative of fake news.

Finally, we see potential in examining user engagement metrics on social media, such as shares and comments, to assess the virality of political content. Analyzing how users interact with news articles can provide valuable insights into which political messages resonate with audiences and why they are shared.

Author Contributions: Conceptualization, Y.B.-F. and A.G.-S.; methodology, Y.B.-F. and J.O.-V.; software, J.O.-V. and J.G.-D.; validation, J.O.-V. and A.G.-S.; formal analysis, Y.B.-F. and A.G.-S.; investigation, Y.B.-F., J.O.-V. and A.G.-S.; data curation, J.O.-V. and J.G.-D.; writing—original draft preparation, Y.B.-F. and A.G.-S.; writing—review and editing, Y.B.-F., A.G.-S. and J.G.-D.; supervision, Y.B.-F. and A.G.-S.; project administration, Y.B.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received support from the Ministry of Science and Innovation of the Government of Spain within the framework of the DEBATrue research project (PDC2021-121720-I00), as well as from the grants for the consolidation and structuring of competitive research units awarded by the Consellería de Educación, Ciencia, Universidades y Formación Profesional of the Xunta de Galicia (reference ED431B 2024/26).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Our Spanish politicial fake news dataset can be downloaded from https://www.kaggle.com/datasets/javieroterovizoso/spanish-political-fake-news. (accessed on 13 September 2024). The fine-tuning and testing codes for the Transformer-based models assessed are available on GitHub at https://github.com/gssi-uvigo/debatrue (accessed on 13 September 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hakak, S.; Alazab, M.; Khan, S.; Gadekallu, T.R.; Maddikunta, P.K.; Khan, W.Z. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Gener. Comput. Syst.* **2021**, *117*, 47–58. [CrossRef]
2. Wah Chu, K.; Xie, R.; Wang, Y. Cross-language fake news detection. *Data Inf. Manag.* **2021**, *5*, 100–109. [CrossRef]
3. Poddar, K.; Bessie, G.; Umadevi, K.S. Comparison of Various Machine Learning Models for Accurate Detection of Fake News. In Proceedings of the 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 22–23 March 2019; Volume 1, pp. 1–5. [CrossRef]
4. Chen, M.Y.; Lai, Y.W.; Lian, J.W. Using Deep Learning models to detect fake news about COVID-19. *ACM Trans. Internet Technol.* **2023**, *23*, 1–23. [CrossRef]
5. Farhangian, F.; Cruz, F.; Cavalcanti, G. Fake news detection: Taxonomy and comparative study. *Inf. Fusion* **2024**, *103*, 102140. [CrossRef]
6. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
7. Rodrawangpai, B.; Daungjaiboon, W. Improving text classification with transformers and layer normalization. *Mach. Learn. Appl.* **2022**, *10*, 100403. [CrossRef]
8. Soyalp, G.; Alar, A.; Ozkanli, K.; Yildiz, B. Improving text classification with Transformer. In Proceedings of the 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 15–17 September 2021; pp. 707–712. [CrossRef]
9. Tezgider, M.; Yildiz, B.; Aydin, G. Text classification using improved bidirectional transformer. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6486. [CrossRef]
10. Shaheen, Z.; Wohlgenannt, G.; Filtz, E. Large scale legal text classification using Transformer models. *arXiv* **2020**, arXiv:2010.12871. [CrossRef]
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
12. Santana, M. Fake News Detector Spanish News Edition. 2023. Available online: https://www.kaggle.com/code/msantrod/fake-news-detector-spanish-news-edition (accessed on 1 November 2023).
13. Morcillo-Montano, J.G.; Flores-Quinayas, J.E. Modelo para la Detección de Noticias Falsas en Formato Texto en la Red Social Twitter, Aplicado al Contexto Político Colombiano de las Elecciones Presidenciales de 2022. Bachor's Thesis, Faculty of Engineering, Master in Data Science, Santiago de Cali, Colombia, 2022.
14. Posadas-Durán, J.P.; Gomez-Adorno, H.; Sidorov, G.; Escobar, J.J. Detection of fake news in a new corpus for the Spanish language. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4868–4876. [CrossRef]
15. Rodríguez-Sánchez, F.; Carrillo-de Albornoz, J.; Plaza, L.; Mendieta-Aragón, A.; Marco-Remón, G.; Makeinenko, M.; Plaza, M.; Gonzalo, J.; Spina, D.; Rosso, P. Overview of EXIST 2022: sEXIsm Identification in Social neTworks. *Soc. Espa Nola Para Proces. Languaje Nat. (NLP)* **2022**, *69*, 229–240.
16. González-Carvajal Centenera, S. Estudio del Rendimiento de BERT Frente a Métodos Clásicos de Procesamiento de Lenguaje Natural. Bachor's Thesis, Informatic Engineering Bachelor, Universidad Autónoma de Madrid, Madrid, Spain, 2020.
17. Alim, A.; Ayman, A.; Kumar Donepudi, P.; Myung, S. Detecting fake news using Machine Learning: A systematic literature review. *arXiv* **2021**, arXiv:abs/2102.04458.
18. Granik, M.; Mesyura, V. Fake news detection using Naïve Bayes classifier. In Proceedings of the IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kyiv, Ukraine, 29 May–2 June 2017; pp. 900–903. [CrossRef]
19. Abdullah-All-Tanvir, M.; Akhter, S.; Huq, M. Detecting fake news using Machine Learning and Deep Learning algorithms. In Proceedings of the 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, 28–30 June 2019; pp. 1–5. [CrossRef]

20. Kudarvalli, H.; Fiaidhi, J. Experiments on detecting fake news using Machine Learning algorithms. *Int. J. Reliab. Inf. Assur.* **2020**, *8*, 15–26. [CrossRef]
21. Vyas, P.; Liu, J.; El-Gayar, O. Fake news detection on the Web: An LSTM-based approach. *Res. Publ.* **2021**, *267*, 1–5.
22. Jain, P.; Sharma, S.; Aggarwal, P. Classifying Fake News Detection Using SVM, Naive Bayes and LSTM. In Proceedings of the 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Virtual, 27–28 January 2022; pp. 460–464. [CrossRef]
23. Nishant, R.; Deepika, K.; Naman, K.; Chandan, R.; Ahad, A. Fake news classification using Transformer based enhanced LSTM and BERT. *Int. J. Cogn. Comput. Eng.* **2022**, *3*, 98–105. [CrossRef]
24. Islam, T.; Hosen, M.; Mony, A.; Hasan, M.; Jahan, I.; Kundu, A. A Proposed Bi-LSTM Method to Fake News Detection. In Proceedings of the International Conference for Advancement in Technology (ICONAT), Goa, India, 21–22 January 2022; pp. 1–5. [CrossRef]
25. Thampi, S.; El-Alfy, E.; Trajkovic, L. Exploring fake news identification using word and sentence embeddings. *J. Intell. Fuzzy Syst.* **2021**, *41*, 5441–5448. [CrossRef]
26. Verma, P.; Agrawal, P.; Amorim, I.; Prodan, R. WELFake: Word embedding over linguistic features for fake news detection. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 881–893. [CrossRef]
27. Truica, C.O.; Apostol, E. It's all in the embedding! Fake news detection using document embeddings. *Mathematics* **2023**, *11*, 508. [CrossRef]
28. Hashmi, E.; Yayilgan, S.; Yamin, M.; Ali, S.; Abomhara, M. Advancing fake news detection: Hybrid deep learning with FastText and explainable AI. *IEEE Access* **2024**, *12*, 44462–44480. [CrossRef]
29. Fang, X.; Wu, H.; Jing, J.; Meng, Y.; Yu, B.; Yu, H.; Zhang, H. NSEP: Early fake news detection via news semantic environment perception. *Inf. Process. Manag.* **2024**, *61*, 103594. [CrossRef]
30. Dhawan, M.; Sharma, S.; Kadam, A.; Sharma, R.; Kumaraguru, P. Game-on: Graph attention network based multimodal fusion for fake news detection. *Soc. Netw. Anal. Min.* **2024**, *14*, 114. [CrossRef]
31. Wu, K.; Yang, S.; Zhu, K.Q. False rumors detection on Sina Weibo by propagation structures. In Proceedings of the IEEE 31st International Conference on Data Engineering, Seoul, Republic of Korea, 13–17 April 2015; pp. 651–662. [CrossRef]
32. Kwon, S.; Cha, M.; Jung, K.; Chen, W.; Wang, Y. Prominent features of rumor propagation in online social media. In Proceedings of the IEEE 13th International Conference on Data Mining, Dallas, TX, USA, 7–10 December 2013; pp. 1103–1108. [CrossRef]
33. Farokhian, M.; Rafe, V.; Veisi, H. Fake news detection using dual BERT deep neural networks. *Multimed. Tools Appl.* **2023**, *83*, 43831–43848. [CrossRef]
34. Tretiakov, A. Fake News in Spanish. 2020. Available online: https://www.kaggle.com/datasets/arseniitretiakov/noticias-falsas-en-espaol (accessed on 1 November 2023).
35. Gómez-Adorno, H.; Posadas-Durán, J.P.; Enguix, G.B.; Porto-Capetillo, C. Overview of FakeDeS at IberLEF 2021: Fake News Detection in Spanish Shared Task. *Proces. Del Leng. Nat.* **2021**, *67*, 223–231.
36. Rodríguez-Sánchez, F.; Carrillo-de Albornoz, J.; Plaza, L. Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access* **2020**, *8*, 219563–219576. [CrossRef]
37. Lakshmipathi, N. IMDB Dataset of 50 K Movie Reviews. 2019. Available online: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews (accessed on 1 November 2023).
38. Howard, A.; Culliton, P.; Guo, Y. Natural Language Processing with Disaster Tweets. 2019. Available online: https://www.kaggle.com/competitions/nlp-getting-started/data (accessed on 1 November 2023).
39. Camargo, F. FASAM—NLP Competition—Predict News Category. 2020. Available online: https://www.kaggle.com/competitions/fasam-nlp-competition-turma-4 (accessed on 1 November 2023).
40. Zhang, C. Tutorial for Chinese Sentiment Analysis with Hotel Review Data. 2017. Available online: https://github.com/Tony607/Chinese_sentiment_analysis/tree/master/data/ChnSentiCorp_htl_ba_6000 (accessed on 1 November 2023).
41. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *Big Data* **2018**, *8*, 171–188. [CrossRef]
42. Martínez-Gallego, K.; Álvarez-Ortiz, A.; Arias-Londoño, J. Fake News Detection in Spanish Using Deep Learning Techniques. *arXiv* **2021**, arXiv:2110.06461.
43. Bugoi, A.; Córdova, D.A.; Corpas, J.; Hernández, D.; Karlgren, E.; Turiel, E. FakeStorage. 2021. Available online: https://github.com/alcorpas10/FakesStorage (accessed on 1 November 2023).
44. Fioretos, D.P. Fake News Detection with the GREEK-BERT Model with a Focus on COVID-19. Bachor's Thesis, National and Kapodistrian University of Athens, Athens, Greece , 2021. Available online: https://pergamos.lib.uoa.gr/uoa/dl/object/2967905/file.pdf (accessed on 1 November 2023).
45. Keya, A.; Wadud, M.A.; Alatiyyah, M.; Hamid, A. AugFake-BERT: Handling Imbalance through Augmentation of Fake News Using BERT to Enhance the Performance of Fake News Classification. *Appl. Sci.* **2022**, *12*, 8398. [CrossRef]
46. Reitz, K. Requests: HTTP for Humans. 2011. Available online: https://docs.python-requests.org/en/latest/ (accessed on 1 November 2023).
47. Richardson, L. Beautiful Soup Documentation. 2004. Available online: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed on 1 November 2023).

48. OpenAI. Official Website. Available online: https://openai.com/ (accessed on 1 November 2023).
49. Honnibal, M.; Montani, I. Spacy Website. 2015. Available online: https://spacy.io (accessed on 1 November 2023).
50. Honnibal, M.; Montani, I. Spacy's NER Model. 2018. Available online: https://spacy.io/models/es#es_core_news_lg (accessed on 1 November 2023).
51. Python Software Foundation. Python Documentation. Available online: https://www.python.org/doc/ (accessed on 1 November 2023)
52. Chouikhi, H.; Chniter, H.; Jarray, F. Arabic Sentiment Analysis Using BERT Model. In *Advances in Computational Collective Intelligence*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 621–632. [CrossRef]
53. Muller, M.; Salathe, M.; Kummervold, P.E. COVID-Twitter-BERT: A Natural Language Processing model to analyse COVID-19 content on Twitter. *Front. Artif. Intell.* **2020**, *6*, 1023281. [CrossRef]
54. Wang, Z.; Ng, P.; Ma, X.; Nallapati, R.; Xiang, B. Multi-Passage BERT: A Globally Normalized BERT Model for Open-Domain Question Answering. *arXiv* **2019**, arXiv:1908.08167
55. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for self-supervised learning of language representations. *arXiv* **2020**, arXiv:1909.11942.
56. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
57. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
58. Casola, S.; Lauriola, I.; Lavelli, A. Pre-trained transformers: An empirical comparison. *Mach. Learn. Appl.* **2022**, *9*, 100334. [CrossRef]
59. Wang, C.; Huang, S.; Zhou, Y. Sentiment analysis of MOOC reviews via ALBERT-BiLSTM model. *MATEC Web Conf.* **2021**, *336*, 05008. [CrossRef]
60. Akpatsa, S.; Lei, H.; Li, X.; Obeng, V.H.; Martey, E.; Addo, P.; Fiawoo, D. Online News Sentiment Classification Using DistilBERT. *J. Quantum Comput.* **2022**, *4*, 1–11. [CrossRef]
61. Adel, H.; Dahou, A.; Mabrouk, A.; Abd, E.; Kayed, M.; El-Henawy, I.M.; Alshathri, S.; Amin-Ali, A. Improving crisis events detection using DistilBERT with hunger games search algorithm. *Mathematics* **2022**, *10*, 447. [CrossRef]
62. Angin, M.; Beyza, T.; Arda, C.; Demiralp, G. Atay, M.; Angin, P.; Dikmener, G. A RoBERTa approach for automated processing of sustainability reports. *Sustainability* **2022**, *14*, 16139. [CrossRef]
63. Akhila, N.; Sanjanasri, J.P.; Soman, K. Comparative study of BERT models and RoBERTa in Transformer based question answering. In Proceedings of the 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 23–25 June 2023; pp. 1–5. [CrossRef]
64. Yuzhe, Z.; Zhang, H. FinBERT-MRC: Financial named entity recognition using BERT under the machine reading comprehension paradigm. *Neural Process. Lett.* **2023**, *55*, 1–21. [CrossRef]
65. Liu, Y.; Wei, S.; Huang, H.; Lai, Q.; Li, M.; Guan, L. Naming entity recognition of citrus pests and diseases based on the BERT-BiLSTM-CRF model. *Expert Syst. Appl.* **2023**, *234*, 121103. [CrossRef]
66. Wu, X.; Xia, Y.; Zhu, J.; Wu, L.; Xie, S.; Qin, T. A study of BERT for context-aware neural machine translation. *Mach. Lang.* **2022**, *111*, 917–935. [CrossRef]
67. Prasanthi, K.; Eswari, M.; Sabarinadh, S.; Naga, D.; Sravani, B. A Novel Approach for Sentiment Analysis on social media using BERT and ROBERTA Transformer-Based Models. In Proceedings of the IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 7–9 April 2023; pp. 1–6. [CrossRef]
68. Talaat, A.S. Sentiment analysis classification system using hybrid BERT models. *J. Big Data* **2023**, *10*, 110. [CrossRef]
69. Jia, K.; Meng, F.; Liang, J.; Gong, P. Text sentiment analysis based on BERT-CBLBGA. *Comput. Electr. Eng.* **2023**, *112*, 109019. [CrossRef]
70. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is multilingual BERT? *arXiv* **2019**, arXiv:1906.01502.
71. Vásquez, J.; Gómez-Adorno, H.; Bel-Enguix, G. BERT-based approach for sentiment analysis of Spanish reviews from TripAdvisor. *Proces. Del Leng. Nat.* **2021**, *67*, 223–231.
72. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish pre-trained BERT model and evaluation data. *arXiv* **2022**, arXiv:2308.02976.
73. Goyal, N.; Du, J.; Ott, M.; Anantharaman, G.; Conneau, A. Larger-scale transformers for multilingual masked language modeling. *arXiv* **2021**, arXiv:2105.00572.
74. de la Rosa, J.; Ponferrada, E.G.; Villegas, P.; Gonzalez, P.; de Prado, S.; Romero, M.; Grandury, M. BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *arXiv* **2022**, arXiv:2207.06814.
75. Crawl, C. Multilingual Common Crawl Corpus (MC4). 2020. Available online: https://huggingface.co/datasets/mc4 (accessed on 1 November 2023).
76. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the North American Chapter of the Association for Computational Linguistics. *arXiv* **2020**, arXiv:2010.11934.

77. Gutiérrez-Fandiño, A.; Armengol-Estapé, J.; Pamies, M.; Llop-Palao, J.; Silveira-Ocampo, J.; Carrino, C.; González-Aguirre, C.; Armentano-Oller, C.; Rodríguez-Penagos, C.; Villegas, M. Spanish language models. *arXiv* **2021**, arXiv:2107.07253.
78. Wolf, T. Hugging Face Transformers. 2019. Available online: https://github.com/huggingface/transformers (accessed on 1 November 2023).
79. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.