mathematics

MDPI

Article

Leveraging Zero and Few-Shot Learning for Enhanced Model Generality in Hate Speech Detection in Spanish and English

José Antonio García-Díaz, Ronghao Pan * and Rafael Valencia-García

Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain; joseantonio.garcia8@um.es (J.A.G.-D.); valencia@um.es (R.V.-G.)
* Correspondence: ronghao.pan@um.es

Abstract: Supervised training has traditionally been the cornerstone of hate speech detection models, but it often falls short when faced with unseen scenarios. Zero and few-shot learning offers an interesting alternative to traditional supervised approaches. In this paper, we explore the advantages of zero and few-shot learning over supervised training, with a particular focus on hate speech detection datasets covering different domains and levels of complexity. We evaluate the generalization capabilities of generative models such as T5, BLOOM, and Llama-2. These models have shown promise in text generation and have demonstrated the ability to learn from limited labeled data. Moreover, by evaluating their performance on both Spanish and English datasets, we gain insight into their cross-lingual applicability and versatility, thus contributing to a broader understanding of generative models in natural language processing. Our results highlight the potential of generative models to bridge the gap between data scarcity and model performance across languages and domains.

Keywords: hate speech detection; zero-shot learning; few-shot learning; fine-tuning; large language models; natural language processing

MSC: 68T50

Citation: García-Díaz, J.A.; Pan, R.; Valencia-García, R. Leveraging Zero and Few-Shot Learning for Enhanced Model Generality in Hate Speech Detection in Spanish and English. *Mathematics* **2023**, *11*, 5004. https://doi.org/10.3390/math11245004

Academic Editor: Florentina Hristea

Received: 9 November 2023
Revised: 11 December 2023
Accepted: 14 December 2023
Published: 18 December 2023

1. Introduction

Online social networks have evolved into vast interconnected communities that function as communication platforms, facilitating the exchange of information and social discourse. While these virtual spaces undoubtedly enhance global connectivity, they also raise a troubling concern: the spread of hate speech. Hate speech encompasses a range of discriminatory and biased behaviors, including homophobia, misogyny, racism, transphobia, and other forms of intolerance, which affect individuals as well as online communities and platforms that strive to create inclusive and safe environments. Identifying and mitigating instances of hate speech on social media platforms is critical to protecting the digital sphere from the harmful effects of prejudice, hostility, and harassment.

In the ongoing fight against hate speech in online spaces, the field of Natural Language Processing (NLP) has evolved significantly in recent years. Traditional methods of hate speech detection, often based on statistical approaches and conventional machine learning classifiers, have been outpaced by advances in deep learning. In particular, Automatic Document Classification (ADC) using Transformers has emerged as the new frontier in the fight against online hate. These powerful models, with their ability to learn complex patterns in language and context, have achieved unprecedented accuracy and efficiency in distinguishing hate speech from benign content. Their success has led to a paradigm shift in how we approach this multifaceted problem.

While Transformers have undoubtedly demonstrated exceptional performance in controlled and simulated environments, their effectiveness has faced notable challenges when applied to the unpredictable and dynamic landscape of real-world online social networks. The discrepancy between idealized laboratory conditions and the complexity of

the online ecosystem has raised concerns about the generalizability of the models. These discrepancies call for a deeper examination of their adaptability to diverse and evolving hate speech contexts. However, the latest approaches in NLP, such as Large Language Models (LLMs), have the ability to directly handle a wide range of NLP tasks and domains, and they possess Zero-Shot Learning (ZSL) and Few-Shot Learning (FSL) capabilities. Thus, the central motivation of this research is to evaluate the potential of ZSL and FSL approaches, which are specifically designed to address the very issue of generalization and adaptability. By subjecting generative models such as BLOOM [1] or LLAMA [2] to a battery of real-world Spanish and English hate speech datasets, we seek to uncover whether these models exhibit improved generalization and robustness in the fight against hate speech compared to traditional fine-tuning approaches.

In this case, the evaluation of the datasets for English and Spanish was chosen because English is the most spoken language and Spanish is the fourth [3], even though both are typologically different languages, one belonging to the Germanic languages and the other to the Romance languages [4].

To evaluate the performance of ZSL and FSL capabilities compared to fine-tuning strategies, we defined the following research questions:

- RQ1. Do ZSL and FSL strategies improve the performance of fine-tuning an LLM for hate speech detection?
- RQ2. Are current ZSL and FSL models equally good at detecting hate speech in English and Spanish?
- RQ3. What are the best generative LLMs for performing ZSL and FSL classification in hate speech detection?
- RQ4. Are the same models equally valid for ZSL and FSL in hate speech detection?

The rest of the manuscript is organized as follows. First, in Section 2 the reader will find the state-of-the-art in hate speech detection and different strategies for performing ZSL and FSL experiments. Next, Section 3 describes the evaluated dataset and the pipeline for performing the comparisons between ZSL and FSL in comparison with fine-tuning approaches. Next, Section 4 presents the results which are evaluated in Section 5. Finally, the conclusions of the paper as well as and promising lines of research can be found in Section 6.

2. State-of-the-Art

Hate speech can be defined as the use of language that promotes discrimination, hostility, or violence against individuals or groups based on their race, ethnicity, religion, gender, sexual orientation, disability, or other protected characteristic [5]. Hate speech can take many forms and is often targeted at specific groups, resulting in types such as racism, xenophobia, homophobia, misogyny, transphobia, and more. These types of hate speech are characterized by their specific prejudices and discriminatory attitudes, highlighting the diversity of groups that may be targeted or marginalized by such expressions. Hate speech is an important social and ethical concern because it can contribute to real harm, perpetuate stereotypes, and undermine inclusivity and tolerance in society.

Hate speech detection has undergone a paradigm shift, driven by the evolution of NLP. Transformer-based models, which are the building blocks of Large Language Models (LLMS), exemplified by BERT, RoBERTa, and their multilingual counterparts, have become the focus of modern hate speech detection systems. Their ability to capture contextual linguistic information has revolutionized the field. In contrast, earlier methods relied on statistical features such as TF–IDF or non-contextual word embeddings such as GloVe [6] or fastText [7].

In a survey published in 2018 in [5], the authors highlighted the lack of hate speech detection systems for non-English languages. Since then, a few datasets have been published on this topic, especially those published in shared tasks in workshops. However, in recent surveys, such as the one published in [8], in which the authors evaluate the most important datasets published in recent years on the topic of hate speech, the authors conclude that

several datasets in the bibliography do not have sufficient examples and are therefore not reliable for hate speech detection. In Spanish, the authors of [9] evaluated which features and which feature integration techniques are most effective for hate speech detection. They focus mainly on transformers and linguistic features, and two strategies for combining the features: knowledge integration and ensemble learning. The evaluation was carried out on four Spanish datasets on different types of hate speech. Two of them were published in workshops as shared tasks. They were the shared tasks (1) AMI 2018 [10], held at IberEval 2018 and which focused on the detection of misogyny; and (2) HatEval 2019 [11], held at SemEval 2019 and which focused on the detection of hate speech against immigrants and women. The other two datasets are (3) the full Spanish MisoCorpus 2020 [12], which focused on misogyny; and (4) HaterNET [13], a binary dataset compiled from Twitter. The authors concluded that the integration of linguistic features with the transformers using the knowledge integration strategy outperformed other approaches in identifying hate speech in Spanish.

Zero and Few-Shot Learning

In recent years, many studies have addressed the problem of so-called low-resource languages and the possibilities of using multilingual approaches based on LLMs. In [14], evidence was found that Multilingual BERT (mBERT), a multilingual masked language model based on transformers, is capable of zero-shot cross-lingual transfer. Furthermore, in [15], the ability of this model to transfer syntactic knowledge between languages was investigated by examining whether and to what extent syntactic dependencies learned in one language are maintained in others. In [16,17], the compressibility of the BERT model was verified, specifically its ability to capture linguistic knowledge in word representations.

In particular, some have focused on the transfer of specific knowledge or phenomena into phylogenetically different languages by ZSL and FSL of LLMs. For example, the authors of [18] explored the problem of multilingual transfer in unseen languages where no unlabeled data are available for pre-training a model. A sentiment analysis task in 12 languages, including 8 unseen languages, was used to analyze the effectiveness of different few-shot learning strategies. Another similar paper [19], where the ability of the pre-trained BERT neural model in Italian to embed syntactic dependency relations in its layers by approximating a dependency parse tree was investigated. For this purpose, a structural probe, a supervised model capable of extracting linguistic structures from a language model, was trained using the contextual embeddings of BERT layers.

Regarding the evaluation of novel ZSL and FSL strategies in deep learning, the work described in [20] measures the reliability of using state-of-the-art generative LLMs to build knowledge graphs. In this sense, the authors propose a novel strategy for asking different LLMs to extract the data to build the knowledge graph. This strategy is based on ZSL, since no requirements are needed to guide the prompts. Another work evaluating ZSL capabilities is [21], in which the authors propose ChatIE, which combines ZSL strategies and ChatGPT for a question-answering task. The evaluated task is divided into several subtasks, including the extraction and recognition of entities and their relations. The authors evaluate a total of six datasets written in two languages. Their proposed model outperforms models trained in the traditional way (i.e., full-shot models).

The paper published in [22] comes closest to our proposal. Among other research objectives, the authors evaluate the ZSL performance of different LLMs and hate speech using the HatEval 2019 dataset [11]. Five LLMs posing as different human annotators are evaluated. While the results are promising, the authors conclude that human annotation is still needed. The main differences with our work are that no few-shot learning capabilities are evaluated and that hate speech is only evaluated in one dataset.

3. Materials and Methods

This section describes the experiments conducted to answer the proposed research questions regarding the performance of ZSL and FSL in detecting hate speech. Therefore,

this section is divided into two parts. The first, Section 3.1, describes the datasets evaluated in our proposal. These datasets are in Spanish and English. Next, Section 3.2 describes the pipeline for carrying out the experiments. This pipeline includes three strategies: fine tuning of an LLS, defined as baseline, and ZSL and FSL.

3.1. Datasets

This section describes the datasets used to evaluate the performance of the ZSL and FSL features. In order to select the datasets that help us answer the RQs defined in this work, we focus on hate speech datasets in two languages: Spanish and English. Another goal is to cover different subtopics of hate speech, such as the detection of sexist or misogynistic content, or racism, transphobia, and homophobia.

In order to make the results comparable across datasets, we focused on a unique task: binary hate speech detection. That is, we select datasets that allow us to identify which texts contain hate speech and which do not. It is worth noting that most of the selected datasets come from shared tasks in workshops that defined a binary classification task. However, there are a few datasets that we have adapted to meet this requirement. Another important point is that not all datasets published in the workshops had the gold labels published. In these cases, we reorganized the dataset to create a new test set from the training split. Therefore, the results in these cases are not comparable to those published in the official task rankings.

The selected datasets are described below, but a summary can be found in Table 1, which includes their publication year, language, hate speech subdomain, and size.

- **EXIST** (EXIST 2021-es, EXIST 2022-es, EXIST 2022-es, EXIST 2022-en): These are a series of shared tasks focused on identifying sexism in Spanish and English. There are editions of EXIST in 2021 [23], 2022 [24], and 2023 [25] in different international workshops such as CLEF or IberLEF. The challenges proposed to the participants usually consist of a binary classification of sexist comments and multi-classification problems to explain why the comments are sexist. In this work, we focus on the binary classification task of 2021 and 2022, with the datasets of Spanish and English separately. The golden labels are not published for these datasets, so we have chosen a custom split for testing in this work.
- **HaterNet 2019** (HaterNet). The HaterNet 2019 dataset [13] contains 6k documents annotated as hateful and non-hateful. The dataset can be accessed at 8 November 2023 (https://zenodo.org/record/2592149#.YNBqJGj7SUl). This dataset is unbalanced, since only about 1.5k documents are annotated as hateful. The original evaluation of the dataset focuses on the F1 score of the hateful class. This dataset has the gold labels of the test split.
- **HatEval 2019** (HatEval). The HatEval [11] shared task took place in SemEval 2019, and is about detecting hate speech against immigrants and women. The dataset is in two languages: Spanish and English, and it was collected from Twitter. In our work, we focus on the first subtask of the competition, which is about binary classification to detect hate speech. This dataset has the gold labels of the test split.
- **Spanish hate speech detection in football** (Football) [26]. In this paper, the authors published a dataset for hate speech detection in Spanish, consisting of almost 7.5k football-related tweets. These tweets were manually categorized as aggressive, racist, misogynist, and safe. In the work, the authors proposed a multi-label approach, and achieved a macro F1 score of 88.713% with the combination of LLM features within the same neural network. This dataset has the gold labels of the test split.
- **Spanish MisoCorpus 2020** (MisoCorpus). The Spanish MisoCorpus 2020 dataset [12] focuses on the binary identification of misogyny. This dataset is almost balanced. It can be downloaded in the full version or divided into three splits regarding different categories. The first one focuses on the violence against relevant women; the second one is about the messages from Spain and Latin America to understand cultural and background differences; and the last one is about general characteristics related to misogyny. This dataset has the gold labels of the test split.

- **Explainable Detection of Online Sexism** [27] (EDOS). This shared task was conducted in SemEval 2023 and focused on detecting and explaining sexism in English. The dataset was collected from Gab and Reddit. In this paper, we focus on the first subtask, binary sexism detection. This dataset has the gold labels of the test split.
- **Hate Speech and Offensive Content Identification in Indo-European Languages, 2020** (HASOC). The HASOC shared task was conducted in FIRE 2020, and it contains documents in English, German, and Hindi for the identification of hateful, offensive and profane content. This dataset has the gold labels of the test split.

It is worth noting that these datasets were selected based on their relation to hate speech, rather than other common datasets for understanding assessment such as GLUE [28]. Furthermore, the selected datasets have been used in international workshops such as IberLEF or CLEF.

Table 1. Year, language, hate speech subdomain, and size of the datasets.

Dataset	Year	Language	Domain	Size
EXIST-2021-es [23]	2021	Spanish	Sexism	3436
EXIST-2022-es [24]	2022	Spanish	Sexism	6233
HaterNet [13]	2019	Spanish	Hate	6000
HatEval [11]	2019	Spanish	Hate	6599
Football [26]	2023	Spanish	Hate	8026
MisoCorpus [12]	2020	Spanish	Misogyny	8390
EXIST-2021-en [23]	2021	English	Sexism	3106
EXIST-2022-en [24]	2022	English	Sexism	6170
HatEval [11]	2019	English	Hate	13,000
EDOS [27]	2022	English	Hate	20,000
HASOC	2020	English	Hate	5124

3.2. Pipeline

3.2.1. Baseline: Fine-Tuning Models

For a fair comparison of the ZSL and FSL capabilities of generative models with fine-tuning LLMs, we established a strong baseline by fine-tuning several popular LLMs based on different architectures (BERT, RoBERTa) and different optimization strategies (distillation) and focusing on a specific dataset or multilingual.

Fine-tuning an LLM for an ADC task involves the process of adapting a model, such as BERT, to a specific classification objective. This is achieved by taking a well-trained LLM and further training it on a labeled dataset containing documents annotated with labels. During this fine-tuning process, the parameters of the LLM are adjusted to learn the patterns and features relevant to the classification task. The goal is to optimize the model's performance in accurately categorizing new documents into predefined labels. Fine-tuning LLMs is a powerful approach that leverages the model's pre-trained language understanding capabilities for ADC tasks such as sentiment analysis, topic categorization, spam detection, and more.

Below is a comparison of the LLMs evaluated.

- **Mono-lingual Transformers**. The two most popular monolingual transformer architectures are BERT (Bidirectional Encoder Representations from Transformers) [29] and RoBERTa (a Robustly Optimized BERT Pre-training Approach) [30]. These models were trained on English data.
  BERT is pre-trained on large amounts of text data to understand the contextual nuances of language. BERT's bidirectional architecture allows it to capture relationships between words and their environment, making it highly effective for various NLP tasks, from sentiment analysis to question answering and more. RoBERTa is an evolution of the original BERT model. It has been trained on a larger and more diverse dataset, using a longer training period and a dynamic masking strategy. Unlike BERT,

RoBERTa does not use the Next Sentence Prediction (NSP) task during pre-training. It also uses a larger vocabulary and incorporates additional training techniques, all of which contribute to its superior performance and robustness in various natural language understanding tasks. Both general-purpose models can be adapted to solve other tasks through a form of transfer learning called fine-tuning. In this process, a pre-trained model is retrained on specific datasets and tasks, and the model's parameters are adjusted to perform well on these new tasks.
There are two LLMs in Spanish, MarIA and BETO. MarIA [31], on the one hand, is trained with the RoBERTa architecture and BETO [32], on the other hand, is trained with the BERT architecture.
We are also evaluating lightweight models: ALBERT [33] and DistilBERT [34]. ALBERT (*A Lite BERT*) is an optimized variant of the BERT model designed to improve computational efficiency without sacrificing performance by significantly reducing the number of parameters. DistilBERT, on the other hand, is a distilled version of the BERT model. It achieves compactness and computational efficiency by using distillation. Distilling involves compressing and simplifying its architecture to create a lighter version while retaining its essential knowledge. The process typically involves training a smaller model (known as the student) to mimic the behavior of the larger, pre-existing model (the teacher). These models have also been adapted to Spanish [35].

- **Multi-lingual Transformers**. Multilingual LLMs are models that have been trained on text from multiple languages, giving them the ability to understand and generate text in different linguistic contexts. Some advantages are that these models facilitate cross-lingual knowledge transfer because they can apply their understanding from one language to another, reducing the need for language-specific models. Second, they are resource efficient, allowing multiple languages to be handled by a single model, thereby reducing computational overhead. In some scenarios, multilingual LLMs require less labeled data to achieve competitive performance on some tasks. However, dedicated monolingual models typically outperform multilingual models.
  In this paper, we evaluate multilingual BERT, one of the first multilingual models, but also two newer models: DeBERTa [36], and TwHIN [37]. DeBERTa stands for Decoding-enhanced BERT with Disentangled Attention. It is a model that improves BERT by enhancing its decoding capabilities and disentangling attention mechanisms, resulting in better performance on various natural language processing tasks. TwHIN is trained on 7 billion microblogging posts from Twitter, making it suitable for short, noisy, and user-generated text often found in hate speech.

To obtain the best result for each dataset and language model, we perform a hyperparameter tuning step to perform the fine-tuning process. For this, we use the RayTune library [38]. This step is as follows. For each dataset and language model, we train a total of 10 models. Each model has different parameters to be evaluated. The hyperparameters are as follows: (1) the training batch size, where 8 or 16 are the only alternatives; (2) the weight decay, with values between 0.0 and 0.3 following a uniform distribution; (3) the warm-up steps, with step values of 0, 250, 500, or 1000; (4) the number of epochs (between 1 and 5); and (5) the learning rate, with values between 1×10^{-5} and 5×10^{-5} following a uniform distribution. The algorithm for selecting the next pair of hyperparameters is based on HyperOptSearch, with the Tree of Parzen Estimators (TPE) and the ASHA scheduler. The goal is to maximize the macro-weighted F1 score.

3.2.2. Generative Models

In terms of text generation models, we have conducted experiments with five state-of-the-art fine-tuned instruction LLMs based mainly on three architectures: (1) T5 with an encoder-–decoder, (2) Llama-2, and (3) BLOOMZ. We specifically chose these five models because they have extensive fine-tuning across a wide range of instructions, making them the most representative of each architecture category. The selected models are described below.

- **Flan-T5**. It is the instruction fine-tuned version of T5 [39] that has achieved strong few-shot performance, even compared to much larger models like PaLM 62B. It has been fine-tuned on over 1000 tasks and covers 60 languages. For this study, we used the XL version of Flan-T5, which contains a total of 3 billion parameters [40].
- **Flan-alpaca**. It is an encoder-–decoder model based on T5 [39] and has been fine-tuned with the Alpaca instruction dataset and GPT4ALL [41].
- **mT0**. It is a model belonging to the BLOOMZ and mT0 family, a group of models capable of understanding human instructions in dozens of languages through zero-shot learning [42]. Specifically, these are fine-tuned models derived from BLOOM and mT5 over a mixture of multilingual tasks. For this paper, we used the large version, which has a total of 1.3 billion parameters.
- **Llama v2**. It is a family of pre-trained LLMs, fine-tuned over a range of 7B to 70B parameters, capable of generating text and summarizing or rewriting existing text [2]. In this case, we used the Stable Beluga 7B and Stable Beluga 13B models, based on Llama-2 with 7B and 13B parameters, fine-tuned with the Orca-style dataset [43]. Note that due to hardware limitations, the Llama-2 13B is loaded with a 4-bit quantization and this fact usually reduces the performance of the model.

3.2.3. ZSL and FSL Prompting

A prompt is a type of input or instruction that is inserted into an LLM to generate a desired response. It can be a sentence, a phrase, or even an entire paragraph, and serves as a starting point or guide for the language model to generate text. Therefore, the proper design and customization of prompts can have a significant impact on the performance of LLMs in specific tasks, such as sentiment analysis.

For ZSL in T5-based models (Flan-T5 and Flan-alpaca), we have defined a prompt in the form of a paragraph consisting mainly of two parts: an instruction to the LLM and the text to be analyzed. In the LLM instruction, to ensure that the models always return one of the classification classes, we introduced a kind of control sequence, as shown in Figure 1. We considered the classification of the aforementioned datasets from a binary perspective. Thus, for the mT0 model, the best performance was achieved with a prompt like "*Is this a sexist tweet?*" and the answer will always be `yes` or `no`. Instructed models of Llama-2 require prompts to be constructed with specific fields: "system", "user", and "assistant". The "system" field is used to specify the instruction or guidance to the system, "user" contains the instance to be classified, and "assistant" is the output indicator.

For the FSL approach, we randomly selected five examples of each label and included them in the prompt using the *Stormtrooper* (https://github.com/centre-for-humanities-computing/stormtrooper/tree/main/stormtrooper(accessed at 8 November 2023)) tool approach, which consists of including the examples in the instruction part of the LLM with the following format: "Please respond with a single label that you think fits the document best. Here are some examples of labels given by experts: examples". The "examples" part is where the randomly extracted examples from the dataset are inserted.

Despite the inclusion of a control sequence in the model, there are still a few cases where the model returns an unrelated response. In these cases, we replaced the response with the most common label in the dataset.

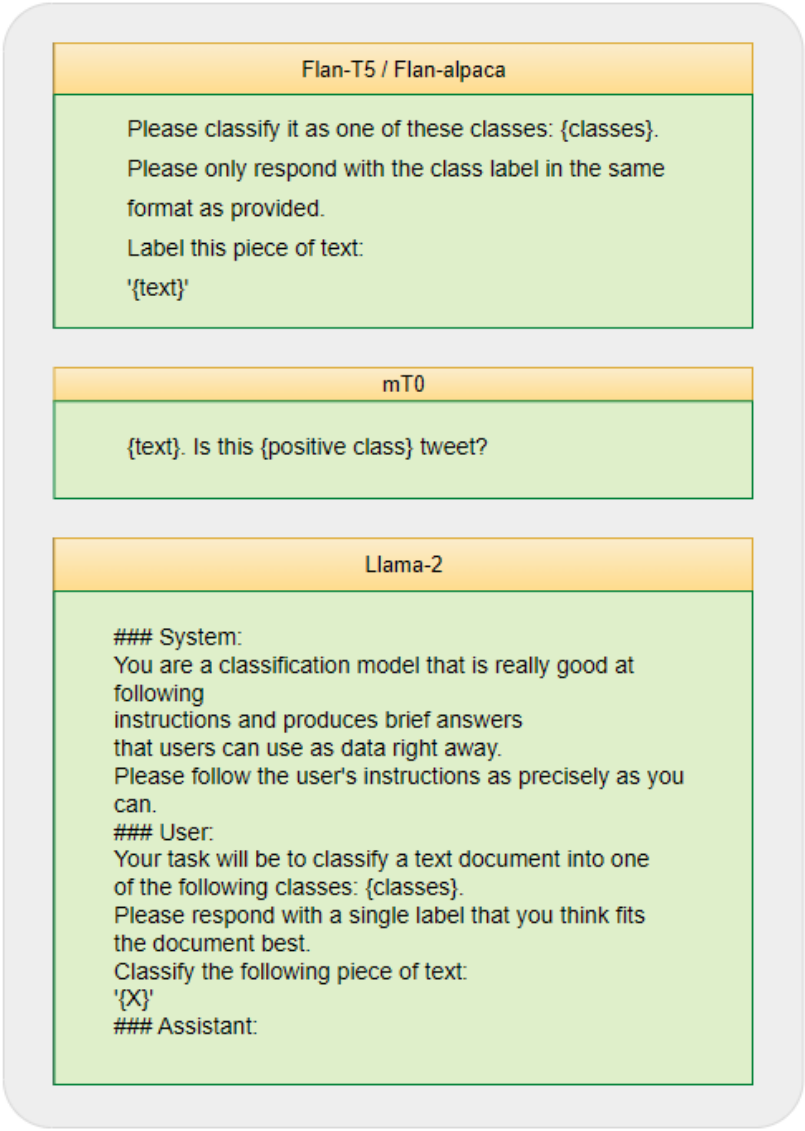

Figure 1. Instructions formulated for ZSL in our study of LLMs for each classification task. The "classes" part indicates the possible labels of the dataset and the "text" part is where the text to be parsed is inserted.

4. Results

In this section, we present the results obtained for the comparison between the fine-tuning and generative models. The results are divided into Spanish (see Section 4.1) and English (see Section 4.2) datasets.

Since we only consider hate speech classification from a binary perspective, the comparison of all models is based only on the hate speech class, including precision, recall, and F1 score. In this sense, we ignore the relevance of the class imbalance between the datasets in our benchmark.

In terms of hardware resources, all experiments are performed on a GeForce RTX 4090 (24 GB). As mentioned earlier, the Llama-2 13B model is evaluated with 4-bit precision due to hardware limitations.

4.1. Spanish Datasets

First, we report the results obtained with the Spanish split of the EXIST dataset in Table 2 for 2021 (left) and 2022 (right) for the positive class (i.e., a document annotated as sexist). Note that this evaluation is performed with a custom validation split, as the gold labels were not released for this shared task. Looking at the results obtained with the fine-tuning strategy, we can see that the two multilingual models, DeBERTa and TwHIN, achieved very good performance on the 2021 dataset. On the other hand, these models obtained more limited results in 2022, where DistilBETO obtained the best F1 score for the sexist label (2022). In this sense, multilingual DeBERTa obtained an almost perfect recall but very limited precision in 2022, which in binary classification indicates that the model is not reliable, as it always predicts that all documents are sexist. It is worth noting that EXIST 2022 is almost twice the size of EXIST 2021. However, monolingual LLMs such as BETO and MarIA give consistent results in both 2021 and 2022, with MarIA slightly better in both cases.

In terms of ZSL, the 7B version of the Llama-2 model achieved the best results in both EXIST 2021 and EXIST 2022 datasets, with F1 scores of 69.883% and 69.872%, respectively. Contrary to the zero-shot scenario, the FSL inference (five shots in our experiments) shows that the performance of Flan-Alpaca, Flan-T5, and 13B Llama-2 did not improve in EXIST 2021 and even worsened due to the introduced examples being poorly correlated with the training data of these models. In the FSL of EXIST 2022, we can see that the five examples selected for each label have improved the performance of Flan-T5, Flan-Alpaca, and Llama-2 13B. The largest absolute gains are obtained with mT0, with an improvement of about 28%.

Table 2. Benchmark of the fine-tuning, zero, and few-shot learning of Spanish datasets of EXIST 2021 (left) and 2022 (right) with the positive class. The results are calculated with a custom validation split. The best results for each metric are shown in bold.

		2021			2022		
	LLM	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Fine-tuning	ALBETO	78.4530	80.2260	79.3296	**79.0484**	75.3927	77.1773
	BETO	80.5882	77.4011	78.9625	77.7597	83.5951	80.5719
	DistilBETO	80.3815	83.3333	81.8308	78.0309	83.6823	**80.7579**
	MarIA	80.5479	83.0508	81.7802	78.0130	83.5951	80.7077
	mBERT	73.2240	75.7062	74.4444	73.4459	83.5078	78.1543
	mDeBERTa	**81.9718**	82.2033	82.0874	50.2413	**99.9127**	66.8613
	TwHIN	78.4615	**86.4406**	**82.2581**	50.0323	67.5393	57.4824
Zero-shot	Flan-T5	**67.1598**	64.1243	65.6069	**64.0981**	63.8743	63.9860
	Flan-alpaca	61.8943	**79.3785**	69.5545	60.8696	79.4066	68.9133
	mT0	63.2653	8.75671	15.3846	58.6538	10.6457	18.0207
	Llama-2	64.8910	75.7062	**69.8827**	55.2178	**95.1134**	**69.8718**
	Llama-2 13B	72.3684	62.1469	66.8693	70.0397	61.6056	65.5525
Few-shot	Flan-T5	**69.8305**	58.1912	63.4823	**67.7686**	64.3979	66.0403
	Flan-alpaca	51.3828	**99.7175**	**67.8194**	53.6176	**97.6440**	**69.2236**
	mT0	51.8868	77.6836	62.2172	52.2752	41.0995	46.0186
	Llama-2	69.2547	62.9944	65.9763	64.1658	56.7190	60.2131
	Llama-2 13B	71.2871	61.0169	65.7534	65.1969	72.2513	68.5430

Next, we evaluate the Spanish split of the HatEval 2019 shared task for discriminating between documents labeled as hateful to immigrants and hateful to women. The results are shown in Table 3. In this case, the performance is obtained with the test set, as the gold labels were released. For the fine-tuning strategy, the best performance for the hateful comments is achieved with DistilBETO, with an F1 score of 76.237%. Looking at the result of the other lightweight model, ALBETO, its performance is also very competitive for detecting hateful comments, with a performance of 75.334%. In general, all the fine-tuned LLMs achieve a similar range of values. The most limited result is obtained with multilingual BERT (70.240%). Nevertheless, the performance of the other multilingual models, mDeBERTA and

TwHIN, is very promising, as they both outperform the monolingual model BETO, although the result of MarIA is slightly better (75.912%). Finally, to compare the performance with the official results of the shared task [11], the overall macro averaged F1 score is 73% and our best macro averaged F1 score (not shown in the table) is 78.45%, also with DistilBETO.

In the ZSL of the hate speech detection models, we can see that Llama-2 from the 7B version achieved the best result with an F1 score of 65.369%, followed by Llama-2 from the 13B version with an F1 score of 64.100%. Regarding the FSL, the examples included in the prompt did not improve the performance of the models. We suspect that this is because the examples have little correlation with the test set, introducing noise into the hate speech prediction. Nevertheless, the 13B version of the Llama-2 model improved its performance by about 2%, achieving an F1 score of 66.283%, surpassing the best ZSL result.

Table 3. Benchmark of the fine-tuning, zero, and few-shot learning of Spanish datasets of HatEval 2019 with the positive class. The results are calculated with the test split. The best results for each metric are shown in bold.

	LLM	Precision	Recall	F1 Score
Fine-tuning	ALBETO	70.2490	81.2121	75.3338
	BETO	66.4216	82.1212	73.4417
	DistilBETO	70.5806	82.8787	**76.2369**
	MarIA	71.2766	81.2121	75.9207
	mBERT	65.6992	75.4545	70.2398
	mDeBERTa	67.2393	**83.0303**	74.3051
	TwHIN	**72.8324**	76.3636	74.5562
Zero-shot	Flan-T5	**65.8228**	55.1515	60.0165
	Flan-alpaca	50.8961	86.0606	63.9640
	mT0	46.3177	79.0909	58.4219
	Llama-2	53.5266	83.9394	**65.3687**
	Llama-2 13B	47.8358	**97.1212**	64.1000
Few-shot	Flan-T5	**74.8428**	36.0606	48.6708
	Flan-alpaca	47.7702	95.7576	63.7418
	mT0	41.2874	96.2121	57.7798
	Llama-2	58.3110	65.9091	61.8777
	Llama-2 13B	53.4323	87.2727	**66.2831**

The next evaluated dataset is the Spanish MisoCorpus 2020, the results of which are shown in the Table 4. This dataset is about misogyny detection with tweets containing hatred towards women with responsibility charges, tweets from different Spanish speaking countries and tweets with different misogynistic characteristics. The strategy of fine-tuning LLMs for the binary classification task yields very high results in terms of precision, recall, and F1 score for the positive label, regardless of the language model. In fact, the difference between the best (mDeBERTa) and the worst (multilingual BERT) is only 1.808% of the F1 score. Regarding ZSL in text generation models for the classification of misogyny texts, we can see that the best result is obtained with the 13B version of Llama-2, with an F1 score of 69.60%. Furthermore, inference with few shots (five shots in our experiments) shows an improvement in all models except mT0. This draws our attention to the large performance loss compared to fine-tuning with ZSL and FSL. Especially in models such as Flan-T5 in ZSL and FSL, or mT0 in FSL, with very limited recall, there is a suggestion that these models give random predictions.

Table 5 shows the results obtained for the detection of hate speech in the football dataset. In this sense, if we observe the results of the fine-tuning strategy, we can see that the best precision and F1 score is obtained with the monolingual model MarIA (87.535% of precision, 85.175% of F1 score), while the multilingual DeBERTa achieved the best recall (85.302%). Multilingual BERT achieved the lowest F1 score (80.926%), but this result is surpassed by another multilingual model, TwHIN, with an F1 score of 83.974%). The lightweight models ALBETO and DistilBETO also achieved very good results, with F1 scores of 84.888% and 84.375%, respectively. This table also shows the performance of different text generation models in a ZSL and FSL scenario. The best result was achieved with the 13B version of

Llama-2, with an F1 score of 72.326% in ZSL. However, we can see that the examples selected for FSL did not improve the performance of the models due to their quality, since FSL models depend heavily on the composition and quality of the test set.

Table 4. Benchmark of the fine-tuning, zero, and few-shot learning of Spanish datasets of Spanish MisoCorpus 2020 with the positive class. The results are calculated with the test split. The best results for each metric are shown in bold.

	LLM	Precision	Recall	F1 Score
Fine-tuning	ALBETO	90.1389	88.5402	89.3324
	BETO	90.3581	89.4952	89.9246
	DistilBETO	89.9587	89.2224	89.5890
	MarIA	89.8649	**90.7230**	90.2919
	mBERT	89.1185	88.2673	88.6909
	mDeBERTa	**90.6849**	90.3138	**90.4990**
	TwHIN	90.6207	89.6316	90.1235
Zero-shot	Flan-T5	**68.0581**	51.1596	58.4112
	Flan-alpaca	51.6817	85.9482	64.5492
	mT0	51.1530	33.2879	40.3306
	Llama-2	51.8270	**94.8158**	67.0203
	Llama-2 13B	57.3959	88.4038	**69.6026**
Few-shot	Flan-T5	**72.0247**	63.5744	67.5362
	Flan-alpaca	46.5176	**99.3179**	63.3594
	mT0	42.6172	83.0832	56.3367
	Llama-2	64.2005	73.3970	68.4914
	Llama-2 13B	62.0619	82.1282	**70.6988**

Table 5. Benchmark of the fine-tuning, zero, and few-shot learning of Spanish datasets of Hate Football Corpus 2023 with the racist class. The results are calculated with the test split. The best results for each metric are shown in bold.

	LLM	Precision	Recall	F1 Score
Fine-tuning	ALBETO	85.0	84.7769	84.8883
	BETO	85.2632	85.0394	85.1511
	DistilBETO	83.7209	85.0394	84.3750
	MarIA	**87.5346**	82.9396	**85.1752**
	mBERT	84.1360	77.9528	80.9264
	mDeBERTa	80.0493	**85.3018**	82.5921
	TwHIN	84.7594	83.2021	83.9735
Zero-shot	Flan-T5	**80.2548**	33.0709	46.8401
	Flan-alpaca	57.8829	67.4541	62.3030
	mT0	48.1061	66.6667	55.8856
	Llama-2	50.9874	74.5407	60.5544
	Llama-2 13B	64.3892	**82.4934**	**72.3256**
Few-shot	Flan-T5	87.3016	14.4357	24.7748
	Flan-alpaca	54.9729	**79.7900**	65.0964
	mT0	26.3636	15.2231	19.3012
	Llama-2	**88.3041**	39.6325	54.7101
	Llama-2 13B	65.5629	78.7798	**71.5663**

Finally, for the Spanish datasets, we report the results of HaterNET 2019 in Table 6. Regarding the fine-tuning strategy, the multilingual model DeBERTa achieved the best performance with an F1 score of 68.858% with the positive (hateful) class. These results outperform the experiments carried out when the dataset was compiled, which had an F1 score of 61.1% [13] based on a neural network combining Long–Short Term Memory (LSTM) and MultiLayer Perceptron (MLP) architectures with features related to words, emoticons, and embeddings enriched with TF–IDF. Similar to other Spanish experiments (see Tables 2 and 3), the most limited results are obtained with multilingual BERT, with an F1 score of 58.519%. In these experiments, we also observed that most models achieve better precision than recall, with the multilingual models DeBERTa and TwHIN being the most notable exceptions. For ZSL on the HaterNET dataset, we can see that the best model

is Llama-2 from the 13B version, which achieved an F1 score of 50.741%. Regarding FSL, we can see that it did not improve the performance of the Flan-T5 and mT0 models due to the fact that the example set is poorly correlated with the training set of these models. However, with the same examples, it improved the performance of Flan-Alpaca and both the 7B and 13B versions of Llama-2, obtaining the best results in FSL with an F1 score of 56.350%, surpassing the best results in ZSL.

Table 6. Benchmark of the fine-tuning, zero, and few-shot learning of Spanish datasets of Spanish HaterNET 2019 with the positive class. The results are calculated with the test split. The best results for each metric are shown in bold.

	LLM	Precision	Recall	F1 Score
Fine-tuning	ALBETO	64.8649	54.3689	59.1549
	BETO	**72.7612**	63.1068	67.5910
	DistilBETO	67.4193	67.6375	67.5283
	MarIA	67.9054	65.0485	66.4463
	mBERT	68.3983	51.1329	58.5185
	mDeBERTa	66.6666	**71.1974**	**68.8576**
	TwHIN	66.0436	68.6084	67.3016
Zero-shot	Flan-T5	**42.0245**	44.3366	43.1496
	Flan-alpaca	33.5535	82.2006	47.6548
	mT0	36.7925	50.4854	42.5648
	Llama-2	30.5328	**96.4401**	46.3813
	Llama-2 13B	35.5383	88.6731	**50.7407**
Few-shot	Flan-T5	**54.0541**	6.4725	11.5607
	Flan-alpaca	36.1613	84.1424	50.5837
	mT0	17.0683	27.5081	21.0657
	Llama-2	42.0382	**85.4369**	**56.3501**
	Llama-2 13B	37.0656	93.2039	53.0387

4.2. English Datasets

In this section, we report the results for the English datasets on the identification of hate speech.

The first experiments use the English splits of the EXIST 2021 and 2022 datasets. The results are shown in the Table 7. Regarding the fine-tuning strategy, BERT is the model that achieves the best results in both datasets, reaching an F1 score of 79.769% in 2021 and 79.682% in 2022. In 2021, BERT also achieves the best precision, but not the best recall, while TwHIN achieves the best precision in 2022. In both cases, the best recall is obtained by the multilingual model DeBERTa, but the low precision obtained indicates that the multilingual DeBERTa always predicts the positive class, making this model useless compared to the others. The lightweight models ALBERT and DistilBERT achieve very competitive results, as well as the multilingual model TwHIN. Looking at the results of ZSL and FSL, we notice that these results are much better than those obtained with the Spanish splits of EXISTS (see Table 2). In fact, Llama-2 (13B) achieves 74.240% of the F1 score in 2021 and 73.962% in 2022 with ZSL. These results are 5.529% below BERT in 2021 and 5.72% in 2022. The performance of FSL is slightly worse in most of the evaluated models, except in the case of mT0.

The next evaluated comparison is with the HASOC 2019 dataset, the results of which are shown in Table 8. Regarding the fine-tuning model strategy, the best performance is achieved by the multilingual model TwHIN, with an F1 score of 86.760% and an almost perfect recall of 93.609%; however, TwHIN is not the model with the best precision, as DistilBERT achieves a precision of 84.754%. All the fine-tuned LLMs achieve similar performance, but as observed with the Spanish datasets (see Section 4.1), the most limited result is obtained with multilingual BERT. From the results obtained in ZSL, we can see that the models perform better in classifying hate speech in English, achieving an F1 score above 70% in all models. Regarding FSL, the performance of Flan-Alpaca has improved, surpassing the best ZSL result with an F1 score of 84.602%.

Table 7. Benchmark of the fine-tuning, zero, and few-shot learning of English datasets of EXIST 2021 (left) and 2022 (right) with the positive class. The results are calculated with a custom validation split. The best results for each metric are shown in bold.

		2021			2022		
	LLM	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Fine-tuning	ALBERT	73.0337	78.7879	75.8017	73.9726	82.1109	77.8296
	BERT	**76.24309**	83.6364	**79.7688**	74.8038	85.2415	**79.6823**
	DistilBERT	74.4505	82.1212	78.0980	74.6479	80.5903	77.5054
	mBERT	70.0831	76.6667	73.2272	73.8731	79.1592	76.4249
	mDeBERTa	48.0349	**100.0**	64.8968	49.4681	**99.8211**	66.1529
	RoBERTa	71.9895	83.3334	77.2472	74.7026	84.2576	79.1929
	TwHIN	73.9612	80.9091	77.2793	**75.6869**	81.3059	78.3959
Zero-shot	Flan-T5	**67.1642**	81.8182	66.7643	**81.5742**	73.4300	73.7705
	Flan-alpaca	61.2159	**88.4848**	72.3668	61.0234	**86.4043**	71.5291
	mT0	55.3672	29.6970	38.6588	61.7094	32.2898	42.3958
	Llama-2	64.3373	80.9091	71.6779	56.2698	95.5277	70.8223
	Llama-2 13B	65.8080	85.1515	**74.2404**	65.3187	85.2415	**73.9620**
Few-shot	Flan-T5	**73.6059**	60.0000	66.1102	**70.4825**	66.6369	68.5057
	Flan-alpaca	54.1176	**97.5758**	69.6216	54.7379	**97.1377**	70.0193
	mT0	47.9279	80.6061	60.1130	49.0061	57.3345	52.8442
	Llama-2	67.1598	68.7879	67.9641	66.5081	62.5224	64.4537
	Llama-2 13B	67.5978	73.3333	**70.3488**	64.2755	80.1431	**71.3376**

Table 8. Benchmark of the fine-tuning, zero, and few-shot learning of English datasets of HASOC 2021 with the positive class. The results are calculated with the test split. The best results for each metric are shown in bold.

	LLM	**Precision**	**Recall**	**F1 Score**
Fine-tuning	ALBERT	81.0872	89.7243	85.1874
	BERT	82.6037	89.8496	86.0744
	DistilBETO	**84.7539**	88.4712	86.5726
	mBERT	80.7474	89.3484	84.8305
	mDeBERTa	82.4074	89.2231	85.6799
	RoBERTa	83.4313	88.9724	86.1128
	TwHIN	80.8442	**93.6090**	**86.7596**
Zero-shot	Flan-T5	**81.1180**	81.8296	81.4722
	Flan-alpaca	74.3665	**95.6140**	**83.6623**
	mT0	64.6825	81.7043	72.2038
	Llama-2	70.6767	94.2356	80.7734
	Llama-2 13B	72.4521	89.9749	80.2683
Few-shot	Flan-T5	**90.2527**	31.3283	46.5116
	Flan-alpaca	76.9231	**93.9850**	**84.6024**
	mT0	59.7484	83.3333	69.5971
	Llama-2	74.2553	87.4687	80.3222
	Llama-2 13B	72.5198	91.6040	80.9524

The results with the EDOS 2023 dataset are shown in Table 9, where monolingual BERT achieves the best performance for the fine-tuning strategy, with an F1 score of 73.795%. It also achieves the best recall (75.773%), but not the best precision, which is achieved by DistilBERT (77.203%). The most limited result is achieved by ALBERT (70.049% of the F1 score), followed by multilingual BERT (70.192% of the F1 score). Compared to BERT, RoBERTa also achieves a good performance with an F1 score of 71.680%, but the multilingual TwHIN surpasses this result with an F1 score of 72.083%. The text generation models for classifying sexist text in the EDOS dataset performed best in the ZSL scenario, with Flan-T5 achieving an F1 score of 53.12%. In the FSL scenario, it improved this result by about 8%, achieving an F1 score of 61.57%.

Table 10 shows the results of HatEval 2019 with the English dataset. Regarding the fine-tuning strategy, the best result is obtained with the multilingual TwHIN, with an F1 score of 67.977% over the positive class. However, the precision of all the LLMs is very

limited for the positive class since the recall is almost perfect in every case. This behavior is not observed in the Spanish part of the HatEval 2019 dataset, where the recall is around 75% and 83%. However, the maximum result obtained in the official ranking for the English dataset was a macro average F1 score of 65.10% [11]. Regarding the ZSL and FSL strategies, the performance of the models is very similar, as almost all models achieve limited precision but high recall, but this suggests that these models also always predict the positive class. However, Llama-2 is the best performer for both ZSL and FSL. Specifically, the best overall result is achieved with Llama-2 for FSL, when the highest overall performance is achieved (F1 score of 67.083%).

Table 9. Benchmark of the fine-tuning, zero, and few-shot learning of English datasets of EDOS 2023 with the positive class. The results are calculated with test split. The best results for each metric are shown in bold.

	LLM	Precision	Recall	F1 Score
Fine-tuning	ALBERT	74.3917	66.1856	70.0491
	BERT	71.9178	**75.7732**	**73.7952**
	DistilBERT	**77.2033**	67.7320	72.1581
	mBERT	72.8381	67.732	70.1923
	mDeBERTa	75.1412	68.5567	71.6981
	RoBERTa	74.2541	69.2783	71.68
	TwHIN	72.8421	71.3402	72.0833
Zero-shot	Flan-T5	**37.2007**	92.8866	**53.1250**
	Flan-alpaca	31.7258	94.9485	47.5600
	mT0	31.1571	49.6907	38.2996
	Llama-2	28.5887	**97.3196**	44.1948
	Llama-2 13B	33.0914	93.8272	48.9270
Few-shot	Flan-T5	**50.1622**	79.6907	**61.5691**
	Flan-alpaca	27.3882	**97.8351**	42.7959
	mT0	24.1716	91.7526	38.2631
	Llama-2	39.9890	74.7423	52.1020
	Llama-2 13B	40.0659	75.1029	52.2548

Table 10. Benchmark of the fine-tuning, zero, and few-shot learning of English dataset of HatEval 2019 with the positive class. The results are calculated with test split. The results are calculated with test split. The best results for each metric are shown in bold.

	LLM	Precision	Recall	F1 Score
Fine-tuning	ALBERT	42.7975	97.6190	59.5065
	BERT	47.0161	97.5397	63.4486
	DistilBERT	45.6329	97.8571	62.2413
	mBERT	45.3933	96.1905	61.6794
	mDeBERTa	45.9650	**98.0952**	62.598
	RoBERTa	46.1831	96.5079	62.4711
	TwHIN	**47.5988**	97.5397	**63.9771**
Zero-shot	Flan-T5	**45.0873**	98.3333	61.8263
	Flan-alpaca	42.8523	99.6825	59.9380
	mT0	44.6973	91.9841	60.1609
	Llama-2	44.8768	**99.7619**	**61.9059**
	Llama-2 13B	44.2918	**99.7619**	61.3470
Few-shot	Flan-T5	50.8132	91.7460	65.4031
	Flan-alpaca	42.1546	1.000000	59.3081
	mT0	42.1414	**97.4603**	58.8404
	Llama-2	**62.3891**	72.5397	**67.0826**
	Llama-2 13B	48.4294	96.6667	64.5298

5. Discussion

Tables 11 and 12 present a comparison showing the best results obtained by different datasets and approaches for the Spanish and English datasets, respectively. In general, we can observe that the fine-tuning approach for transformer models in classification has achieved better performance than ZSL and FSL, but at a higher computational cost.

These results answer RQ1, which asks whether zero and few-shot improve the results of fine-tuning for hate speech detection. In the ZSL approach to hate speech classification in Spanish, the models achieved competent results even though they were not explicitly trained for it, as in the case of the fine-tuning approach. The best model for ZSL was Llama-2 in its 7B and 13B versions.

Regarding FSL, we experimented with a prompt-based FSL using five random examples for each label, and we inserted them into the prompts of the text generation models to guide the model towards better performance. However, based on the results obtained, we can see that the FSL approach did not improve the performance of ZSL, and this is largely due to the quality of the selected few-shot dataset and its relationship with the pre-trained data of the models. Furthermore, finding a set of examples that generalize the concept of hate speech is quite challenging [44]. In this paper [45], an additional retrieval module based on sentence transformers was used to maximize the few-shot performance in clinical and biomedical tasks. However, there are still cases where few-shot learning has worsened the performance of ZSL. Therefore, it would be convenient to select the examples using some kind of heuristic or a method to search for phrases that are more related to a certain class.

If we compare the results obtained for the Spanish and English datasets, we can see that the results obtained by the three strategies evaluated (fine-tuning, ZSL, FSL) are more similar for the English datasets, but greater for the Spanish ones. For example, in EXISTS 2021, there is a 12.402% decrease in performance between the fine-tuning and ZSL strategies in Spanish. However, this difference is only 5.529% in English. Moreover, if we look at the results comparing monolingual and multilingual approaches to fine-tuning, we see that there is a tie in Spanish, as DistilBETO and MarIA are the best performing models in three datasets, while TwHIN and DeBERTa, two multilingual LLMS, achieve the best results in the other three Spanish datasets. In the case of the English datasets, English BERT performed best in both EXISTS 2021 and 2022 and in EDOS, and TwHIN performed best in HatEval and HASOC. In the case of ZSL and FSL, all evaluated models are multilingual. It was therefore expected that the difference in performance would be the same in both languages. Since the results show the opposite, we answer RQ2 (are current ZSL and FSL models equally good at detecting hate speech in English and Spanish?) that ZSL and FSL are better at detecting hate speech in English than in Spanish. However, this comparison must be made with caution, as English and Spanish are typologically different languages with different roots.

With regard to RQ3, which asks about the best generative LLMs for performing ZSL and FSL classification in hate speech detection, we observed that Llama-2 13B is the model that obtained a better result in five of the evaluated datasets for ZSL: three Spanish and two English. In the case of Spanish, the other evaluated version of Llama achieved the best performance in the rest of the evaluated datasets and only one other dataset in English. For the rest of the evaluated English datasets, Flan T5 and Alpaca performed best for EDOS and HASOC. In the case of FSL, Llama-2 13B also achieved the best results in three of the Spanish datasets (HatEval, Football and MisoCorpus), tying with ZSL in two of them (Football and MisoCorpus). Flan-alpaca achieved the best results for the two Spanish EXIST datasets, and Llama-2 for HaterNET. In the case of English, the same models that performed best on ZSL also performed best on FSL. This behavior was not observed for the Spanish datasets. Given these results, we can conclude that Llama-2 13B is the best performing model for zero and few-shot classification in hate speech detection, but this model is not a silver bullet, as there are six datasets where this model did not achieve the best results.

Finally, RQ4 asks whether the same generative LLMs are equally good for zero and few shots. The results show that only two of the Spanish datasets agree (Llama-2 13B in soccer and the MisoCorpus). In English, however, the same models are the best for both ZSL and FSL. So, in this case, the results suggest that the answer to this RQ4 is that it depends on the language. However, if we look at the results individually across all the datasets and generative models evaluated, the difference between ZSL and FSL is usually small, with ZSL performing better. There are exceptions. For example, mT0 shows a difference

of 46.832% between FSL and ZSL in the Spanish EXIST 2021 dataset and a difference of 27.998% in 2022 (see Table 2). In other cases, there are strong differences between ZSL and FSL, both in Spanish and in English. This fact suggests that experiments are needed to evaluate which strategy is better depending on the dataset.

Table 11. Resume of the results of fine-tuning, zero, and few-shot learning for the Spanish datasets.

	Fine-Tuning		ZSL		FSL	
Dataset	**F1 Score**	**Model**	**F1 Score**	**Model**	**F1 Score**	**Model**
EXIST-2021-es	82.2581	TwHIN	69.8827	Llama-2	67.8194	Flan-alpaca
EXIST-2021-es	80.7579	DistilBETO	69.8718	Llama-2	69.2236	Flan-alpaca
HatEval	76.2369	DistilBETO	65.3687	Llama-2	66.2831	Llama-2 13B
HaterNET	68.8576	mDeBERTa	50.7407	Llama-2 13B	56.3501	Llama-2
Football	85.1752	MarIA	72.3256	Llama-2 13B	71.5663	Llama-2 13B
MisoCorpus	90.4990	mDeBERTa	69.6026	Llama-2 13B	70.6988	Llama-2 13B

Table 12. Resume of the results of fine-tuning, zero, and few-shot learning for the English datasets.

	Fine-Tuning		ZSL		FSL	
Dataset	**F1 Score**	**Model**	**F1 Score**	**Model**	**F1 Score**	**Model**
EXIST-2021-en	79.7688	BERT	74.2404	Llama-2 13B	70.3488	Llama-2 13B
EXIST-2022-en	79.6823	BERT	73.9620	Llama-2 13B	71.3376	Llama-2 13B
HatEval	63.9771	TwHIN	61.9059	Llama-2	67.0826	Llama-2
EDOS	73.7952	BERT	53.1250	Flan-T5	61.5691	Flan-T5
HASOC	86.7596	TwHIN	83.6623	Flan-alpaca	84.6024	Flan-alpaca

6. Conclusions and Outlook

In this research, we compare and contrast different strategies for detecting hate speech. In particular, we evaluate two alternatives based on prompting, known as zero and few-shot, against a fine-tuning strategy. Our main goal is to test the generalization ability of these models to detect hate speech in texts written in English or Spanish. Through rigorous evaluation on diverse hate speech detection datasets spanning different domains and languages, we uncovered key insights. The evaluation highlighted the robust generalization capabilities of generative models such as T5, BLOOMZ, and Llama-2, underscoring their potential to bridge the gap between data scarcity and model performance. However, the results are still more limited in performance compared to fine-tuning strategies, but with less time and hardware resources. Our research not only contributes to the evolving landscape of hate speech detection, but also underscores the ability of generative models to advance the fight against online intolerance and discrimination.

In order to unravel the potential of zero and few-shot learning strategies in the field of hate speech detection, a number of core research questions were defined. First and foremost, we investigated the impact of these strategies on fine-tuning language models (LLMs) to improve performance (RQ1). In addition, our research ventured into the cross-lingual landscape by investigating whether these strategies are equally effective for hate speech detection in English and Spanish (RQ2). We delved into the intricacies of generative LLMs to identify the best models for zero and few-shot classification in hate speech detection (RQ3). Finally, we questioned the versatility of these models by exploring whether they are equally valid in the context of zero- and few-shot learning for hate speech detection (RQ4). Our research efforts have been driven by these questions and have provided valuable insights into the evolving field of hate speech detection strategies.

The results show that the performance of models based on T5, BLOOMZ, and Llama-2 is still more limited than the fine-tuning of an LLM for hate speech detection, but the results are more stable with English datasets compared to Spanish. The results also show the potential of Llama-2 13B, which achieved the best performance in most of the datasets. Moreover, we observe a large variability in terms of precision and recall, which suggests that a deep experimentation is still needed for each case to determine which is the best performing model to perform ZSL and FSL. Another interesting finding is that FSL strategies

usually do not outperform ZSL. These results may be due to a poor selection of examples used as input to the FSL models.

These results also suggest that the selection of the best strategy for hate speech detection is highly dependent on the dataset and the model. Therefore, further research should be conducted to find the similarities and differences of the evaluated linguistic models and strategies. In this sense, we propose to combine the use of linguistic features [46] and explicable machine learning tools, such as SHAP and LIME, [47] to analyze the results across datasets. In particular, we propose to compare the results in similar datasets, such as those of EXIST, which published a Spanish and an English variant in the same competition.

As a promising line of research, we propose to build a retrieval module based on Sentence Transformers to identify the subset that generalizes the concept of hate speech from the training set. The idea would be to fine-tune a Sentence Transformers model through contrastive learning [48] for extracting examples for prompt-based FSL, thus maximizing its performance. In this sense, we also propose to improve the quality of the prompts used and to evaluate different strategies for selecting the examples for FSL. Another line we propose is the use of hyperparameter optimization for text generation models. It is also worth noting that, due to hardware limitations, the 7B version of the Llama-2 model was loaded into the GPU with 8-bit precision, and the 13B version with 4-bit precision. In this sense, the comparison between the two models is unfair (although Llama v2 achieved better performance in most experiments). Therefore, we recommend evaluating both models with 8-bit and 4-bit precision.

Finally, we will also propose to evaluate FSL and ZSL capabilities in other domains. We propose two domains. The first one is author profiling, where the number of publications per author is quite large, so the capabilities of ZSL and FSL models will imply a large time saving of resources if the results have the same performance. In this sense, we will evaluate the generative models with the dataset published in [49], which contains demographic and psychographic traits of politicians and journalists from Spain. The second domain is subjective language. Therefore, we will evaluate these models with the Spanish SatiCorpus 2021 [50], which contains pairs of satirical and real digital news, in order to check which models are better suited to discriminate between them. We also propose to evaluate standard reference datasets for model evaluation, such as GLUE [28] and those similar.

Author Contributions: Conceptualization, J.A.G.-D. and R.V.-G.; data curation, R.P.; funding acquisition, R.V.-G.; investigation, R.P.; project administration, R.V.-G.; resources, R.V.-G.; software, J.A.G.-D. and R.P.; supervision, R.V.-G.; visualization, J.A.G.-D.; writing—original draft, all. All authors have read and agreed to the published version of the manuscript.

Funding: This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

Data Availability Statement: Source code for training the zero and few-shot models is available at https://github.com/NLP-UMUTeam/mathematics-zsl-fsl-hate-speech (accessed on 8 November 2023). No new data are created in this research. Therefore it is necessary to request the datasets from the original authors of each paper evaluated in this work.

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

References

1. Scao, T.L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; Gallé, M.; et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv* **2022**, arXiv:2211.05100.
2. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.
3. Cong Khanh, L. English as a Global Language: An Exploration of EFL Learners' Beliefs in Vietnam. *Int. J. TESOL Educ.* **2022**, *3*, 19–33. [CrossRef]
4. Nichols, J. *Linguistic Diversity in Space and Time*; University of Chicago Press: Chicago, IL, USA, 2018.
5. Fortuna, P.; Nunes, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv. CSUR* **2018**, *51*, 1–30. [CrossRef]
6. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.

7. Mikolov, T.; Grave, É.; Bojanowski, P.; Puhrsch, C.; Joulin, A. Advances in Pre-Training Distributed Word Representations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
8. Alkomah, F.; Ma, X. A literature review of textual hate speech detection methods and datasets. *Information* **2022**, *13*, 273. [CrossRef]
9. García-Díaz, J.A.; Jiménez-Zafra, S.M.; García-Cumbreras, M.A.; Valencia-García, R. Evaluating feature combination strategies for hate speech detection in spanish using linguistic features and transformers. *Complex Intell. Syst.* **2023**, *9*, 2893–2914. [CrossRef]
10. Fersini, E.; Rosso, P.; Anzovino, M. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. *IberEval SEPLN* **2018**, *2150*, 214–228.
11. Basile, V.; Bosco, C.; Fersini, E.; Debora, N.; Patti, V.; Pardo, F.M.R.; Rosso, P.; Sanguinetti, M. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MI, USA, 6–7 June 2019; pp. 54–63.
12. García-Díaz, J.A.; Cánovas-García, M.; Colomo-Palacios, R.; Valencia-García, R. Detecting misogyny in Spanish tweets: An approach based on linguistics features and word embeddings. *Future Gener. Comput. Syst.* **2021**, *114*, 506–518. [CrossRef]
13. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and monitoring hate speech in Twitter. *Sensors* **2019**, *19*, 4654. [CrossRef]
14. Chi, E.A.; Hewitt, J.; Manning, C.D. Finding Universal Grammatical Relations in Multilingual BERT. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 6–10 July 2020; pp. 5564–5577. [CrossRef]
15. Guarasci, R.; Silvestri, S.; De Pietro, G.; Fujita, H.; Esposito, M. BERT syntactic transfer: A computational experiment on Italian, French and English languages. *Comput. Speech Lang.* **2022**, *71*, 101261. [CrossRef]
16. Jawahar, G.; Sagot, B.; Seddah, D. What Does BERT Learn about the Structure of Language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Florence, Italy, 28 July–2 August 2019; pp. 3651–3657. [CrossRef]
17. Hewitt, J.; Manning, C.D. A Structural Probe for Finding Syntax in Word Representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MI, USA, 2–7 June 2019; Volume 1: Long and Short Papers, pp. 4129–4138. [CrossRef]
18. Winata, G.; Wu, S.; Kulkarni, M.; Solorio, T.; Preotiuc-Pietro, D. Cross-lingual Few-Shot Learning on Unseen Languages. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Virtual, 20–23 November 2022; Volume 1: Long Papers, pp. 777–791.
19. Guarasci, R.; Silvestri, S.; De Pietro, G.; Fujita, H.; Esposito, M. Assessing BERT's ability to learn Italian syntax: A study on null-subject and agreement phenomena. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *14*, 1–15. [CrossRef]
20. Carta, S.; Giuliani, A.; Piano, L.; Podda, A.S.; Pompianu, L.; Tiddia, S.G. Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction. *arXiv* **2023**, arXiv:2307.01128.
21. Wei, X.; Cui, X.; Cheng, N.; Wang, X.; Zhang, X.; Huang, S.; Xie, P.; Xu, J.; Chen, Y.; Zhang, M.; et al. Zero-shot information extraction via chatting with chatgpt. *arXiv* **2023**, arXiv:2302.10205.
22. Plaza-del Arco, F.M.; Nozza, D.; Hovy, D. Leveraging Label Variation in Large Language Models for Zero-Shot Text Classification. *arXiv* **2023**, arXiv:2307.12973.
23. Rodríguez-Sánchez, F.; Carrillo-de Albornoz, J.; Plaza, L.; Gonzalo, J.; Rosso, P.; Comet, M.; Donoso, T. Overview of exist 2021: Sexism identification in social networks. *Proces. Leng. Nat.* **2021**, *67*, 195–207.
24. Rodríguez-Sánchez, F.; Carrillo-de Albornoz, J.; Plaza, L.; Mendieta-Aragón, A.; Marco-Remón, G.; Makeienko, M.; Plaza, M.; Gonzalo, J.; Spina, D.; Rosso, P. Overview of exist 2022: Sexism identification in social networks. *Proces. Leng. Nat.* **2022**, *69*, 229–240.
25. Plaza, L.; Carrillo-de Albornoz, J.; Morante, R.; Amigó, E.; Gonzalo, J.; Spina, D.; Rosso, P. Overview of exist 2023–learning with disagreement for sexism identification and characterization. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Thessaloniki, Greece, 18–21 September 2023; pp. 316–342.
26. Montesinos-Cánovas, E.; Garcia-Sánchez, F.; Garcia-Díaz, J.A.; Alcaraz-Mármol, G.; Valencia-García-Sánchez, R. Spanish hate speech detection in football. *Proces. Leng. Nat.* **2023**, *71*, 15–27.
27. Kirk, H.; Yin, W.; Vidgen, B.; Röttger, P. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Toronto, ON, Canada, 10–31 January 2023; pp. 2193–2210. [CrossRef]
28. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; Brussels, Belgium, 1 November 2018; pp. 353–355. [CrossRef]
29. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K.N. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MI, USA, 2–7 June 2019; pp. 4171–4186.
30. 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.
31. Gutiérrez Fandiño, A.; Armengol Estapé, J.; Pàmies, M.; Llop Palao, J.; Silveira Ocampo, J.; Pio Carrino, C.; Armentano Oller, C.; Rodriguez Penagos, C.; Gonzalez Agirre, A.; Villegas, M. MarIA: Spanish Language Models. *Proces. Leng. Nat.* **2022**, *68*, 1–22.

32. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish Pre-Trained BERT Model and Evaluation Data. In Proceedings of the PML4DC at ICLR 2020, Addis Ababa, Ethiopia, 26 April 2020.
33. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
34. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
35. Cañete, J.; Donoso, S.; Bravo-Marquez, F.; Carvallo, A.; Araujo, V. ALBETO and DistilBETO: Lightweight Spanish Language Models. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 4291–4298.
36. He, P.; Gao, J.; Chen, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv* **2021**, arXiv:2111.09543
37. El-Kishky, A.; Markovich, T.; Park, S.; Verma, C.; Kim, B.; Eskander, R.; Malkov, Y.; Portman, F.; Samaniego, S.; Xiao, Y.; et al. Twhin: Embedding the twitter heterogeneous information network for personalized recommendation. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 2842–2850.
38. Liaw, R.; Liang, E.; Nishihara, R.; Moritz, P.; Gonzalez, J.E.; Stoica, I. Tune: A research platform for distributed model selection and training. *arXiv* **2018**, arXiv:1807.05118.
39. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
40. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling instruction-finetuned language models. *arXiv* **2022**, arXiv:2210.11416.
41. Chia, Y.K.; Hong, P.; Bing, L.; Poria, S. INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models. *arXiv* **2023**, arXiv:2306.04757.
42. Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T.L.; Bari, M.S.; Shen, S.; Yong, Z.X.; Schoelkopf, H.; et al. Crosslingual generalization through multitask finetuning. *arXiv* **2022**, arXiv:2211.01786.
43. Mukherjee, S.; Mitra, A.; Jawahar, G.; Agarwal, S.; Palangi, H.; Awadallah, A. Orca: Progressive Learning from Complex Explanation Traces of GPT-4. *arXiv* **2023**, arXiv:2306.02707.
44. Mozafari, M.; Farahbakhsh, R.; Crespi, N. Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning. *IEEE Access* **2022**, *10*, 14880–14896. [CrossRef]
45. Labrak, Y.; Rouvier, M.; Dufour, R. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. *arXiv* **2023**, arXiv:2307.12114.
46. García-Díaz, J.A.; Vivancos-Vicente, P.J.; Almela, A. Umutextstats: A linguistic feature extraction tool for spanish. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 6035–6044.
47. Nguyen, H.T.T.; Cao, H.Q.; Nguyen, K.V.T.; Pham, N.D.K. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In Proceedings of the FPT AI Conference, Ha Noi, Vietnam, 6–7 May 2021; pp. 1–6.
48. Gunel, B.; Du, J.; Conneau, A.; Stoyanov, V. Supervised Contrastive Learning for Pre-Trained Language Model Fine-Tuning. *arXiv* **2020**, arXiv:2011.01403.
49. García-Díaz, J.A.; Jiménez-Zafra, S.M.; Valdivia, M.T.M.; García-Sánchez, F.; Ureña-López, L.A.; Valencia-García, R. Overview of PoliticEs 2022: Spanish Author Profiling for Political Ideology. *Proces. Leng. Nat.* **2022**, *69*, 265–272.
50. García-Díaz, J.A.; Valencia-García, R. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex Intell. Syst.* **2022**, *8*, 1723–1736. [CrossRef]