Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2024.049631

ARTICLE

Comparing Fine-Tuning, Zero and Few-Shot Strategies with Large Language Models in Hate Speech Detection in English

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

Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo, Murcia, 30100, Spain

*Corresponding Author: José Antonio García-Díaz. Email: joseantonio.garcia8@um.es

Received: 12 January 2024 Accepted: 02 April 2024 Published: 08 July 2024

ABSTRACT

Large Language Models (LLMs) are increasingly demonstrating their ability to understand natural language and solve complex tasks, especially through text generation. One of the relevant capabilities is contextual learning, which involves the ability to receive instructions in natural language or task demonstrations to generate expected outputs for test instances without the need for additional training or gradient updates. In recent years, the popularity of social networking has provided a medium through which some users can engage in offensive and harmful online behavior. In this study, we investigate the ability of different LLMs, ranging from zero-shot and few-shot learning to fine-tuning. Our experiments show that LLMs can identify sexist and hateful online texts using zero-shot and few-shot approaches through information retrieval. Furthermore, it is found that the encoder-decoder model called Zephyr achieves the best results with the fine-tuning approach, scoring 86.811% on the Explainable Detection of Online Sexism (EDOS) test-set and 57.453% on the Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (HatEval) test-set. Finally, it is confirmed that the evaluated models perform well in hate text detection, as they beat the best result in the HatEval task leaderboard. The error analysis shows that contextual learning had difficulty distinguishing between types of hate speech and figurative language. However, the fine-tuned approach tends to produce many false positives.

KEYWORDS

Hate speech detection; zero-shot; few-shot; fine-tuning; natural language processing

1 Introduction

Large Language Models (LLMs) are gaining popularity in academic and industrial circles due to their exceptional performance in numerous applications. This success is due to their ability to generate human-like text using huge training datasets and billions of parameters. LLMs have versatile capabilities, unlike previous models that were limited to specific tasks. Due to their remarkable performance in various applications, ranging from general Natural Language Processing (NLP) tasks to specific domain functions, they are increasingly being adopted by various domains, including healthcare, education, law, finance, and scientific research [1].

Supervised machine learning consists of training the model on a labeled dataset. However, labeling the dataset is one of the tedious and intensive tasks in NLP. With the emergence of GPT-3 [2], the

in-context learning approach is introduced. In-context learning uses Zero-Shot Learning (ZSL) and Few-Shot Learning (FSL) to teach LLMs to understand the task using natural language text. ZSL enables model training without task-specific data, using prior knowledge for unseen tasks. Conversely, FSL trains models with minimal task examples, facilitating adaptation to new tasks with limited data. Both methods increase model flexibility and generalization capabilities. Therefore, LLMs can be used to directly predict labels or specific classification tasks through prompting. In [3–5], this hypothesis has been substantiated and has shown significant efficacy across domains and classification tasks, in some cases approaching the performance of fine-tuned systems.

In recent years, the Internet and social media have revolutionized the way people communicate. Online platforms allow users to express their opinions freely, as the anonymity provided allows people to feel more free in how they express themselves. However, anonymity has also created a negative side, as it is easy to post false statements and opinions about facts without being challenged, which has encouraged hate speech and offensive content on social media. For this reason, there has been a growing interest in the study of hate speech and online sexism in recent years. Online sexism is a pervasive and harmful phenomenon that can harm targeted women, make online spaces inaccessible and unwelcoming, and perpetuate social asymmetries and injustices [6]. Online Hate Speech (HS) is defined as any communication that disparages an individual or group based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics [7].

The following Research Questions (RQ) are proposed in this paper:

- **RQ1:** Can LLMs identify sexism and hate speech using ZSL and FSL?
- **RQ2:** Which generative models are better at identifying sexism and hate speech?
- **RQ3:** Does the fine-tuning approach of an encoder-decoder or encoder-only model trained by a set of instructions work better than ZSL and FSL for a classification task?

To this end, we evaluate 4 state-of-the-art LLMs for detecting sexist speech and hate speech online using the Explainable Detection of Online Sexism (EDOS) and Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (HatEval) datasets, respectively. For this evaluation, different approaches are tested: (1) fine-tuning of different encoder models; (2) fine-tuning of LLMs; (3) prompting the models in a ZSL and FSL phase with instructions similar to those we would give to a human annotator; (4) a retrieval model based on Sentence Transformers to find texts more related to each class to see if it would improve the FSL performance of the models. Thus, our contribution is to evaluate the efficiency of different LLMs in detecting online hate speech and sexist discourse based on contextual learning. In addition, a comparison is made with other deep learning approaches to hate speech and sexism detection, such as fine-tuning pre-trained language models.

The rest of the paper is organized as follows. Section 2 presents background information on LLMs, including transformer language models and ZSL and FSL approaches for classifying offensive online texts. Next, Section 3 describes the system architecture and the dataset used for evaluation, providing a comprehensive overview of all components and formats. Section 4 details the methodology and procedures used to conduct the experiments of the two approaches. Section 5 covers the evaluation of the results and the error analysis performed for the overall system evaluation. Finally, Section 7 presents the conclusions and outlines future work.

2 State of the Art

This paper evaluates several approaches to classifying online sexism and hate speech, including fine-tuning encoder-only and encoder-decoder models, prompting the LLMs in ZSL and FSL stages

with instructions, and using a retrieval module based on Sentence Transformers to improve FSL performance. This section introduces the concept of LLMs (see Section 2.1), the notion of ZSL and FSL (see Section 2.2), and the state of the art in online offensive speech detection (see Section 2.3).

2.1 Large Language Models

In general, LLMs refer to Transformer-based language models with hundreds of billions or more parameters, trained on large text corpora. Examples of LLMs are GPT-3 [2], LLaMa [8], BART [9], BLOOM [10] or T5 [11] among others. LLMs have demonstrated the ability to understand natural language and to solve complex tasks through text generation.

Currently, LLMs are primarily based on the Transformers architecture [12], which is characterized by multiple stacked layers of attention mechanisms within a deep neural network. These models use similar pre-training goals, such as language modeling, as used in BERT [13] and RoBERTa [14]. However, LLMs significantly increase model size, data volume, and computational resources, as research shows that scalability significantly improves the performance of LLMs [2].

Unlike other language models, LLMs presents a set of unique skills that can be applied to solve a variety of tasks [1].

- **In-Context Learning.** Encoder-decoder LLMs can generate expected outputs for test instances without the need for additional training or gradient updates. This can be done by using natural language instructions or task demonstrations. Thus, LLMs have the capacity for ZSL and FSL.
- **Instruction Following.** When LLMs are fine-tuned with multitask datasets formatted with natural language descriptions (referred to as instruction tuning), they demonstrate the ability to excel at unseen tasks described in a similar way [15]. This allows LLMs to follow task instructions for new tasks without explicit examples, improving their generalization abilities.
- **Step-by-Step Reasoning.** LLMs can handle complex tasks that require multiple steps of reasoning using the chain-of-thought (CoT) prompting strategy [16].

As a result of these, LLMs can be applied in several domains [1]. In healthcare, ChatGPT have shown effectiveness in extracting biological information [17], consulting medical advice [18], analyzing mental health [19], and simplifying reports [20]. In education, LLMs have shown the potential to perform at student level on standardized tests in subjects such as physics and computer science [21], and to act as writing or reading assistants [22]. In the field of law, LLMs have analyzed legal documents [23], predicted judgments [24], and written legal opinions [25]. They have also demonstrated strong legal interpretation and reasoning skills [26], with models such as GPT-4 achieving high scores on simulated bar exams. Legal prompt engineering enhances their comprehension and reasoning skills. In finance, LLMs have been used in financial sentiment analysis [27] or named entity recognition [28], highlighting the existence of specific finance-focused LLMs such as FinGPT [19] and BloombergGPT [29]. In the scientific research, LLMs have also proven beneficial at various stages of scientific research, assisting with literature review, hypothesis generation, data analysis, and scientific writing [1].

However, LLMs face some significant challenges. In healthcare, for example, there are concerns about misinformation and privacy due to potential misinterpretation of medical terms and sharing of sensitive health data. In education, questions remain about plagiarism, bias in AI-generated content, over-reliance on LLMs, and equitable access for non-English speakers. In law, legal challenges such as copyright, privacy, bias, and discrimination remain a concern. In addition, LLMs require careful oversight due to the potential risk of misleading financial content. And in the area of scholarly

research, despite their usefulness in automating review and supporting the research pipeline, concerns remain about the quality of scholarly content generated and the potential for misleading information.

2.2 Zero-Shot and Few-Shot Learning

The richness of representation and the ability to generalize allow advanced LLMs to perform new tasks reasonably well without additional training, known as ZSL [2]. For example, LLMs trained on multilingual texts can perform translations without specific training for that task. ZSL rely solely on instructions in prompts, without any training data, unlike models that have been fine-tuned on thousands of annotated examples. The difference between FSL and ZSL is that the prompts use a limited number of examples for these tasks.

Novel LLMs are sequence-to-sequence Transformers, where both the input and the output are text [11]. When used for classification, the text to be classified is input to the model, and the model outputs a text-based label. In addition, LLMs can be provided with textual prompts that provide additional context and instructions alongside the input text. Therefore, an emerging field known as *prompt engineering* explores the effectiveness of different prompting strategies for text generation and interaction with other generative models.

ZSL predicts a class that the model did not see during training. ZSL can be thought of as transfer learning, since it uses a model trained for one task in a different application than the one for which it was originally trained. In ZSL, the model is given a prompt or string of text that describes the instructions to the model. ZSL excludes all examples of the desired task, but FSL includes one or a few examples of the new task. The effectiveness of a model in a ZSL or FSL task appears to scale with the size of the model, meaning that larger models (with more trainable parameters or layers) generally perform better. In [3], the ZSL approach was evaluated using 5 state-of-the-art LLMs for 5 classification tasks in different domains and in 4 different languages (English, French, German, and Spanish). It was observed that the performance of the models varies significantly between tasks and labels.

2.3 Online Offensive Speech

The exponential growth in the use of social media has provided a platform for the display of harmful and offensive online behavior, which is on the rise. Offensive behavior on social media is having a significant impact on the mental health of many people. In a few cases, it has become so severe that it has led to the extreme of suicide. Therefore, it is important to identify and eliminate offensive behavior as early as possible to make online platforms safer and more secure [30]. For this reason, in recent times, numerous research studies and shares tasks in NLP have emerged regarding the identification of different types of offensive texts.

Various online offensive behavior detection techniques in social media can be classified into four main types: content-based, sentiment and emotion-based, user profile-based, and network-based, according to the type of features that the detection techniques address [30].

The content-based detection approach focuses on identifying offensive behavior by analyzing the content and context of the text. It relies on explicitly identifying offensive content using keywords, lexicon, and obscene language, using various textual, semantic, syntactic, morphological, typographic, stylometric, pragmatic, and word embedding techniques. For example, in [31], a deep learning-based method was proposed that combines the back-translation method with the paraphrasing technique, using the Transformer model and a mixture of experts to generate paraphrases and augment data for identifying hate speech. In [32], a semi-supervised multilevel neural method for multi-label sexism

classification was proposed. They combined Bidirectional Long Short-Term Memory (Bi-LSTM) and an attention mechanism with the BERT model to effectively train a multi-label sexism classifier.

The sentiment and emotion-based detection approach is based on recognizing offensive content by considering the emotions and sentiments conveyed in the text, such as positive, negative, or neutral. Typically, this approach combines sentiment and emotion lexicons, emoticons, and content-based features (such as Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF–IDF), keywords) to increase the efficiency of identifying offensive behavior, such as [33], which proposed a a framework based on shared sentiment knowledge to identify hate speech. In contrast, in [34], they addressed the issue of multi-class classification of textual hate speech by exploring different text mining features and different machine learning models.

Profile-based detection combines various author profile-based traits such as age, gender, sexual orientation, race, followers, followed accounts, likes and shares, along with user behavior and vocabulary on social networks to detect offensive behavior. For example, in [35], the authors proposed a framework to improve the ability to discriminate between bullying and non-bullying cases by capturing and incorporating temporal patterns into cyberbullying detection models. Meanwhile, in [36], they proposed a deep neural network model that captures the semantic and sociocultural context to identify hate speech and provides an interpretable understanding of the classification decision made by the model.

The network-based detection approach uses graph embeddings and neural network features that take into account the user's relationships with other users, in addition to content-based and user-based features, to identify offensive behavior in online social networks. For example, in [37], they proposed a graph-based method for modeling user interaction sessions, focusing on temporal dynamics and topic coherence to improve cyberbullying detection. In addition, Khan et al. [38] proposed a robust framework for identifying incitement to violence in Urdu tweets, given the neglect in this area. The framework uses language models and one-dimensional convolutional neural networks (1D-CNN) on a *corpus* of Urdu tweets. Several models are compared and it is found that the 1D-CNN with word unigram shows superior performance, outperforming comparable models with an accuracy of 89.84% and a macro F1 score of 89.80%.

In recent years, in different international conferences and evaluation forums, such as Conference and Labs of the Evaluation Forum (CLEF), Iberian Languages Evaluation Forum (IberLEF) and Semantic Evaluation (SemEval), new tasks have emerged that focus on the detection of offensive texts. For example, (1) SemEval-2019 Task 5 [7], a multilingual task aimed at detecting hate speech against immigrants and women on Twitter (HatEval); (2) SemEval-2023 Task 10 [6], divided into three subtasks aimed at classifying online sexist speech; (3) HOMO-MEX in IberLEF 2023 [39], which aims to identify online hate speech against the LGBTQ+ population in Spanish-speaking Mexico; and (4) sEXism Identification in Social neTworks (EXIST) in IberLEF 2021 [40] and CLEF 2023 [41], which aims to detect sexism in a broad sense, from explicit misogyny to subtle expressions implying implicit sexist behavior, among others. Most approaches to this problem rely on fine-tuning various encoder-only language models based on Transformers such as BERT, RoBERTa, and others to identify offensive text in social networks.

Table 1 provides a summary of the aforementioned studies on online offensive speech detection. It can be observed that most of them use a supervised approach with different features and deep-learning approaches, but there is a lack of research related to online offensive speech detection based on context learning and fine-tuning of LLMs. Therefore, in this paper, the datasets EDOS [6] and HatEval [7] have been evaluated using different approaches, including fine-tuning different language models based on

Transformers, and finally we compared with the ZSL and FSL approaches based on LLMs for sexism and hate speech detection.

Table 1: Summary of studies related to online offensive speech

Reference	Domain	Approach	Algorithms
[31]	Hate speech	Supervised	LSTM and CNN
[32]	Sexism	Semi-Supervised	Bi-LSTM with BERT
[33]	Hate speech	Supervised	A neural network based on sentiment knowledge sharing
[34]	Hate speech	Supervised	Text mining features with traditional machine learning models
[35]	Cyberbullying	Supervised	Machine learning models
[36]	Hate speech	Supervised	Bidirectional Gated Recurrent Unit (BiGRU) + Char-grams + Attention, BiGRU + Char-grams + Attention + Feed forward layer
[37]	Cyberbullying	Supervised	Graph based method
[38]	Violence tweets	Supervised	1D-Convolutional Neural Network

3 System Architecture

This section explains the different models and approaches that have been evaluated. These approaches can be divided into two main groups.

In approach 1 (see Fig. 1), the fine-tuning of different LLMs is done by adding a sequence classification layer. This layer consists of a final dense layer with as many neurons as there are output classes for the recognition task. The result is a classification model is obtained for each model.

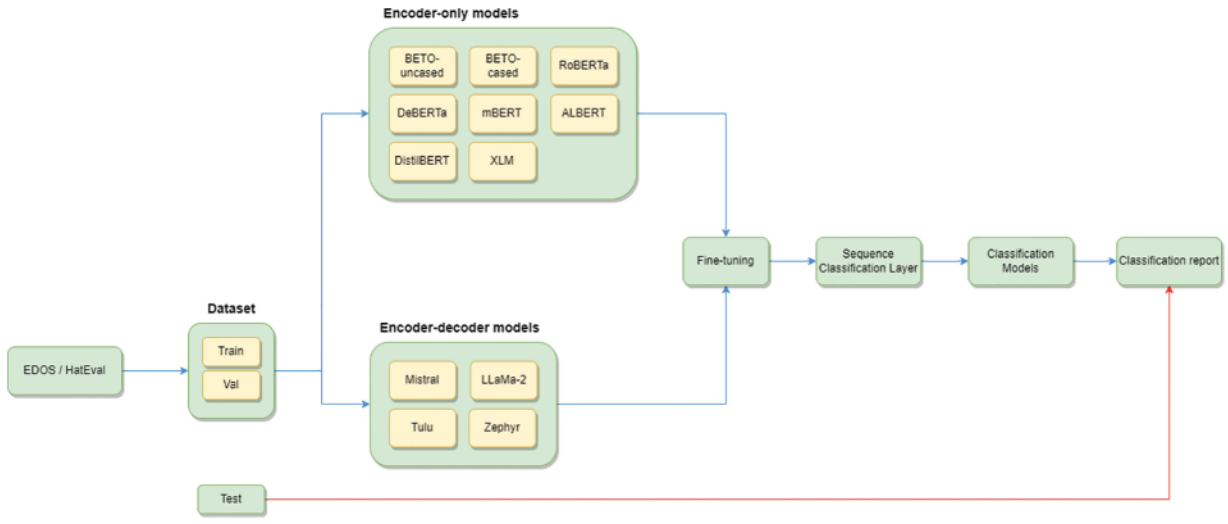

Figure 1: Overall architecture of the approach 1 for the system

In approach 2 (see Fig. 2), the LLMs are used with prompts. This is based on three strategies: (1) ZSL, where the model receives a prompt or text string describing the instructions for the model; (2) 5-shot classification, where the model is given not only the instructions and the text to predict, but also 5 random examples for each output class (for example, for EDOS, 5 examples of sexist texts and 5 of non-sexist texts); and (3) 5-shot classification with Sentence Transformers. This is similar to the previous strategy, but instead of random examples, a Sentence Transformers model is used as a retrieval module to obtain more representative examples of each type.

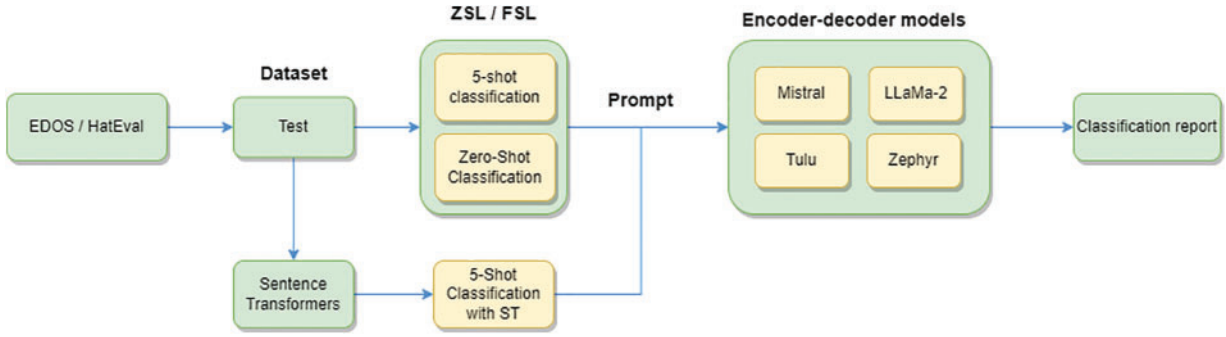

Figure 2: Overall architecture of the approach 2 for the system

3.1 Datasets

For the online sexism detection task, we used the dataset provided by Task 3 of SemEval 2023, known as the *Explainable Detection of Online Sexism (EDOS)*. This dataset is notable for its data diversity, annotation quality, and granularity. The data is collected from two major social media platforms, Reddit and Gab, through a variety of filtering methods, and annotated by highly trained annotators who self-identify as women. Table 2 illustrates the distribution of the dataset, showing that the dataset is divided into three subsets with a ratio of 70-10-20: training, validation, and test sets. The training and validation sets are used for model training and hyperparameter tuning, while the test set is reserved for evaluating the final performance of the model.

Table 2: EDOS dataset distribution

	Training	Validation	Test	Total
Sexist	3,398	486	970	4,854
Not sexist	10,602	1,514	3,030	15,146
Total	**14,000**	**2,000**	**4,000**	**20,000**

Hate speech is characterized by communications that denigrate individuals or groups based on traits such as race, gender, religion, and others. To evaluate different techniques for detecting online hate speech, we used the dataset provided by Task 5 of SemEval 2019 *Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter*. The dataset was collected between July and September 2018, mainly from Twitter, with a focus on content targeting women and immigrants. Various collection strategies were used, including monitoring potential victims and downloading the history of accounts identified as spreading hate, neutral and derogatory keywords, and polarizing

hashtags. The HatEval dataset consists of 13,000 English-language tweets, which we divided into three subsets (training, validation, and test) with a ratio of 70-10-20, as shown in Table 3.

Table 3: HatEval dataset distribution

	Training	Validation	Test	Total
Hate speech	3,783	427	1,740	5,950
Non hate speech	5,217	573	1,260	7,050
Total	**9,000**	**1,000**	**3,000**	**13,000**

3.2 Models

Encoder-only models use only the encoder of a Transformers architecture, using the output generated by it as an input sequence. These models are typically characterized by "bidirectional" attention and are commonly referred to as auto-encoding models [42]. The pre-training of encoder-only models typically involves altering a given sentence in some way, for example by masking random words in it, and asking the model to reconstruct the original sentence. This is known as Masked Language Modeling (MLM). Encoder-only models have shown remarkable performance on tasks that require understanding the entire sentence, such as sentence classification and named entity recognition.

In this study, the following encoder-only models are evaluated, each of them having different architectures and pre-training dataset.

- **BERT.** It is an English bidirectional transformer pre-trained with a combination of MLM and Next Sentence Prediction (NSP) on a large *corpus* consisting of the Toronto Book *corpus* and Wikipedia. It comes in two versions: uncased, which ignores capitalization and removes accents, and cased, which preserves capitalization [13].
- **RoBERTa.** It is a Transformers model pre-trained on a large English *corpus* using an MLM objective. The training *corpus* consists of raw, unlabeled text. Specifically, the model randomly masks 15% of the words in the input, runs the entire masked sentence through the model, and predicts the masked words [14].
- **DeBERTa.** It is an improved version of BERT and RoBERTa that uses disentangled attention and an improved mask decoder. With these two improvements, the model has surpassed RoBERTa's performance on a majority of NLU tasks using 80 GB of training data [43].
- **mBERT.** This is a version of BERT pre-trained on a multilingual *corpus* of 104 languages, mostly from Wikipedia.
- **ALBERT.** It is a lightweight version of BERT that uses parameter reduction techniques to reduce memory consumption and increase the training speed of BERT. In addition, it employs a self-supervised loss that focuses on modeling inter-sentence coherence to consistently support downstream tasks with multi-sentence input [44].
- **DistilBERT.** It is a distilled version of BERT, faster and smaller than BERT. It has been self-trained on the same *corpus* using the BERT base model as a the teacher [45].
- **XLM-RoBERTa.** This is a multilingual model based on RoBERTa, pre-trained with a 2.5 TB filtered CommonCrawl dataset containing 100 languages, using an MLM goal, where it randomly masks 15% of the words in the input, runs the entire masked sentence through the model, and predicts the masked words [46].

Besides, 4 state-of-the-art instruction fine-tuned LLMs from two different model families were evaluated: (1) Mistral-7B-v0.1 [47] and (2) LlaMa-2 [8]. Both families are based on multilayer Transformers-based architecture with encoder and decoder components for text generation. In addition, both families are trained on a large *corpus* of text data and are designed to generate coherent and contextually relevant text. We specifically selected these four models because they have been fine-tuned for a variety of instructions and use intuitive explanations to respond to natural language prompts. This aspect is important when using prompt-based methods (ZSL and FSL). Furthermore, all selected models are available for download as open source software via the HuggingFace hub[1]. In particular, the following encoder-decoder models have been used:

- **Mistral-7B-Instruct.** Mistral-7B is a 7 billion parameter language model designed for superior performance and efficiency. It outperformed the 13 trillion parameter LlaMa-2 model in all benchmarks. This model uses Grouped-Query Attention (GQA) for faster inference and Sliding Window Attention (SWA) to handle sequences of any length with reduced inference cost. The Mistral-7B-Instruct version is a model fine-tuned to follow instructions that has outperformed the LlaMa-2 13B chat model on both human and automated benchmarks [47].
- **Zephyr-7b-beta.** Zephyr-7B-beta [48] is a fine-tuned version of Mistral-7B trained on a combination of synthetic and publicly available datasets using Direct Preference Optimization (DPO).
- **StableBeluga-7B.** It is an auto-regressive language model based on Transformer and fine-tuned on LlaMa-2-7B [8] using a instruction dataset named Orca style dataset [49].
- **Tulu-2.** Open resources for instruction tuning have evolved rapidly, from better base models to new fine-tuning techniques. Tulu-2 [50] is a fine-tuned version of LlaMa-2 with a dataset called Tulu-v2-mix, which is an improved collection of high-quality instruction datasets. In this case, the DPO version of Tulu-2 was used, which consists of a Tulu-2 model trained using the Direct Preference Optimization technique.

It is worth noting the possibility that the LLMs may generate hate speech or sexist discourse, as discussed in [51]. The rapid development of LLMs therefore underscores the importance of ethical considerations and data integrity in AI development, with an emphasis on the FAIR (Findable, Accessible, Interoperable, Reusable) principles. For example, recent generative text models, such as GPT-4 and LlaMa-2, have restrictions on generating text about certain topics that are not ethically appropriate, such as hate, racism, sexism, and so on. The evaluated LLMs are fine-tuned by a set of instructions, which gives us the possibility to control the behavior of the model by a control sequence to the system or by giving instructions to the system at the beginning of the prompt.

4 Experiment Setup

In this section, we detail the methodology and procedures used to conduct the experiments of the two approaches.

4.1 Approach 1: Fine-Tuning

As mentioned in Section 1, one of our goals is to compare the performance of LLMs in classifying hate speech and sexist text, from ZSL or FSL based on prompts to fine-tuning with training and evaluation data.

Unlike traditional supervised machine learning approaches, Transformer-based language models have transfer learning capabilities. This means that they can adapt to new tasks without training the

[1] https://huggingface.co/.

models from scratch, because they transfer the knowledge already acquired during the pre-training phase. Model fine-tuning is a training technique in which the hyperparameters and architecture of the model are modified to fit the training dataset, with the goal of improving the model's performance on a given task. The fine-tuning process is critical to the training of many modern LLMs and can significantly improve performance across multiple tasks [52].

In the first approach, we fine-tune both encoder-only and encoder-decoder type models mentioned in Section 3.2 for sequence classification. That is, we have adapted these models to identify sexist and hate speech texts as a binary classification.

The fine-tuning process can be divided into three main steps: (1) Tokenizing texts using the associated tokenizer of each LLM to convert text strings into integer token IDs that can be read as input to the models. (2) Adding a sequence classification layer on top of the pre-trained models using the *AutoModels* class from the *Transformers* library. (3) Finally, training the models using the training and evaluation sets. For fine-tuning, an epoch-based strategy was used, where a pre-trained model is trained over a certain number of epochs with other hyperparameters fixed. In machine learning, an epoch is a complete iteration through the entire training data set and is validated with the validation set. In this way, after fine-tuning, the strategy saves the model from a particular epoch that has achieved the best result on the validation set. The hyperparameters used are 6 *epochs*, a *learning rate* of 2e-5, a *weight decay* of 0.01, and a *training batch size* of 16 for the encoder-only type models.

It is important to note that large encoder-decoder models such as Mistral, Zephyr, StableBeluga, and Tulu-2 are large models with 7 billion parameters. In contrast, encoder-only models are relatively smaller, making them useful as a basis for comparative studies. Therefore, we use LoRA (Low-Rank Adaptation) [53] for the encoder-decoder models. LoRA accelerates fine-tuning and uses less memory. This method is based on representing weight updates with two smaller matrices (called *update matrices*) through low-rank decomposition. These new matrices can be trained to adapt to new data while keeping the total number of changes low. The hyperparameters used for LoRA were an *r* of 16, a *alpha* of 32, a *dropout* reduction of 0.05, and the same training hyperparameters, with the only change being that the *training batch size* was changed to 1. Finally, we used the logits of the fine-tuned models for the predictions. Table 4 gives a summary of the hyperparameters used for fine-tuning in the experimental setup.

Table 4: The hyperparameters used for fine-tuning of encoder-only and encoder-decoder models

Dataset	Training				Lora		
	Epochs	Learning rate	Weight decay	Batch size	*r*	Alpha	Dropout
Encoder-only	6	2e-5	0.01	16	–	–	–
Encoder-decoder	6	2e-5	0.01	1	16	32	0.05

4.2 Approach 2: Context Learning

A prompt is an input provided to instructed models to guide them on the task they should perform, even without being explicitly trained on that specific task. In a ZSL scenario, a prompt consists of a set of instructions for the model to follow, along with the text to be analyzed. In a FSL scenario, examples are also added so that the model can learn from them to improve its performance.

Fig. 3 shows the prompt instructions that we give to the LLMs that act as our classification model. These instructions are designed for the LLMs to respond with "yes" or "no", as this would facilitate

processing to extract the LLMs' response. However, each model has its own input or instruction format, specifying special tokens to indicate the beginning and end of an instruction.

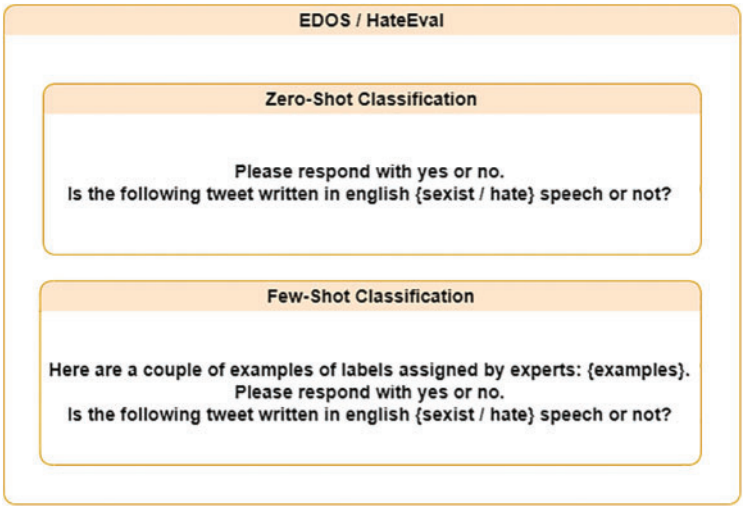

Figure 3: Instruction formulated in our study to prompt the LLMs for each classification task

Mistral-7B-instruct requires that the prompt be surrounded by [INST] and [/INST] tokens to indicate the beginning and end of an instruction. In addition, the first instruction must begin with a sentence start identifier (<s>), and the model itself will generate the response, ending with the sentence end identifier (</s>). For example, the prompt for the hate speech classification task is as follows: <s> [INST] Please respond with yes or no. Is the following tweet written in English hate speech or not? I hate you [/INST].

In the case of Zephyr-7b, prompts must be constructed with specific fields: "System", "User", and "Assistant". The "System" field is used to specify instructions or guidance to the model. The "User" field contains the user's intent and the instance to be classified, while the "Assistant" field is the output indicator. For example, the prompt for the hate speech classification task would be: <|system|>Please respond only with yes or no.</s> <|user|> Is the following tweet written in English hate speech or not? I hate you</s> <|assistant|>.

As for StableBeluga-7B, it uses the same fields as Zephyr-7b, but defined as follows: "### System:", "### User:", and "### Assistant:".

The DPO fine-tuned version of the Tulu model (Tulu-7b-dpo) requires that the model input contain two fields: "user" and "assistant". "user" is used to specify the instructions and the instance to be classified by the model, while "Assistant" is the output indicator. Note that a new line must be inserted after each field, as this can significantly affect the quality of the generation. For instance, the prompt for the hate speech classification task is the following: <|user|> Please respond only with yes or no. Is the following tweet written in English hate speech or not? I hate you</s> <|assistant|>.

Unlike ZSL, a set of examples is inserted into the prompt in FSL, so that the model can learn from them and thus improve its performance. As shown in Fig. 3, the prompt contains a field called "examples", into which the set of examples for each label is inserted. In this study, we evaluated the 5-shot classification, i.e., 5 random examples of each label (in this case, the top 5 in each category in the training set) were inserted into the prompt. We also tested the 5-shot classification approach, using a retrieval module based on Sentence Transformers to retrieve the 5 most semantically similar examples

from the training set for each label. To achieve this, we defined a positive and a negative sentence for each task. For example, for EDOS, the positive sentence would be "This speech is sexist" and the negative sentence would be "This speech is feminist". We then computed the cosine distance between these defined phrases and all the texts present in the vector space to obtain the 5 closest examples of each type.

5 Results and Discussion

In the overview of the EDOS and HatEval tasks, the macro-average F1 score metric is used as a reference. Therefore, in our study, we use this metric to compare the performance of different approaches on the test set of each task.

5.1 Results of Approach 1: Fine-Tuning

Table 5 shows the results of the supervised approach on the dataset, where both the encoder-only and encoder-decoder models are fine-tuned using the training and evaluation sets of the dataset, and the tuned models are then evaluated on the test set. In the case of EDOS, the DeBERTa model achieved the best result among the encoder-only models, with an M-F1 of 83.913. It is also noteworthy that in this scenario, lighter models such as ALBERT and DistilBERT performed better than multilingual models such as mBERT and XLM-RoBERTa, while also requiring less training time. In general, among the encoder-only models, those trained monolingually, specifically on an English *corpus*, outperformed the multilingual models. In the category of encoder-decoder models, Zephyr achieved the highest score with an M-F1 of 86.811, beating DeBERTa by 2.898%. It is also worth noting that Mistral-based models (Mistral-7B and Zephyr) achieved better results than LlaMa-2-based models such as S.Beluga and Tulu.

Table 5: Results of approach 1 for sexist and hate speech classification using the EDOS and HatEval datasets. The F1 score metric of the sexism or hate speech class, the macro F1 score (M-F1) and the weighted F1 score (W-F1) are shown. In addition, metrics such as Macro Precision (M-P) and Macro Recall (M-R) are displayed

Fine-tuning	Model	EDOS				HatEval			
		M-P	M-R	M-F1	W-F1	M-P	M-R	M-F1	W-F1
Encoder models	BETO-uncased	82.942	81.320	82.078	86.988	69.843	59.944	50.121	47.905
	BETO-cased	82.855	79.724	81.093	86.419	67.401	59.024	49.435	47.237
	RoBERTa	82.173	81.580	81.869	86.739	.68.444	57.667	46.281	43.654
	DeBERTa	**84.046**	**83.782**	**83.913**	**88.202**	**70.232**	**61.609**	**53.016**	**51.122**
	mBERT	79.224	77.433	78.252	84.255	64.261	54.505	41.082	37.914
	ALBERT	80.164	79.258	79.692	85.188	65.062	56.947	46.285	43.784
	DistilBERT	81.394	80.588	80.977	86.110	68.536	58.474	47.843	45.403
	XLM	79.863	77.646	78.641	84.584	67.309	57.557	46.519	43.960
Encoder-decoder models	Mistral	85.978	86.104	86.041	89.734	69.717	60.805	51.773	49.756
	S.Beluga	78.655	83.558	80.408	84.891	70.015	63.599	56.677	55.217
	Tulu	81.790	85.060	83.175	87.293	68.756	63.068	56.397	54.979
	Zephyr	**86.969**	**86.656**	**86.811**	**90.330**	**70.845**	**64.279**	**57.453**	**56.035**

Regarding the HatEval dataset, the DeBERTa model achieved the best result among the encoder-only models, with an M-F1 of 53.016. It is observed that lighter models such as ALBERT and DistilBERT outperform certain more complex models such as mBERT and RoBERTa, while also requiring less training time. Overall, it is clear that monolingual encoder-only models outperform multilingual models in hate speech classification. In the category of generative models (encoder-decoder), the fine-tuned Zephyr model achieves the best result with an M-F1 of 57.453, outperforming DeBERTa by 4.437%. It is also observed that LlaMa-based models, such as S.Beluga and Tulu, perform worse than Mistral-based models in the fine-tuning approach.

5.2 Results of Approach 2: Context Learning

Table 6 shows the results of the zero-shot, 5-shot, and 5-shot with Sentence Transformers approaches using the EDOS and HatEval test sets. When using generative models trained on instructional data, the model sometimes fails to classify a text as sexist or hateful and attempts to provide an explanation. To compare the results, we replaced in these cases with "no sexism" or "no hate". This way we can see if there is really is an improvement in the prediction of offensive texts. Therefore, Table 6 shows the F1 scores for the "sexism" and "hate speech" classes in addition to the reference metrics such as the macro-average and the weighted average F1 scores.

Table 6: Results of approach 2 based on the zero-shot, 5-shot, and 5-shot sentence transformer with the encoder-decoder models for sexist and hate speech classification using the EDOS and HatEval datasets. The F1 score metric of the sexism or hate speech class, the macro F1 score (M-F1) and the weighted F1 score (W-F1) are shown

Approach	Metrics	EDOS				HatEval			
		Mistral	S.Beluga	Tulu	Zephyr	Mistral	S.Beluga	Tulu	Zephyr
Zero-shot	Sexism/Hate F1	42.977	44.128	39.301	**54.629**	59.346	64.707	61.519	53.017
	M-F1	45.485	39.874	21.144	**66.304**	55.779	54.461	45.750	**63.813**
	W-F1	46.776	37.683	11.794	**72.317**	55.208	52.822	43.227	**65.540**
5-shot	Sexism/Hate F1	44.042	56.003	46.133	**58.162**	61.023	**67.464**	48.787	63.760
	M-F1	53.878	66.434	49.146	**70.936**	58.740	62.065	56.951	**70.115**
	W-F1	58.943	71.805	50.698	**77.514**	58.375	61.201	58.258	**71.132**
5-shot ST	Sexism/Hate F1	46.802	51.798	51.577	**58.740**	60.325	**64.094**	43.804	58.587
	M-F1	53.335	57.643	63.489	**70.544**	60.924	65.230	55.411	**66.993**
	W-F1	56.700	60.654	69.623	**76.624**	61.020	65.411	57.269	**68.338**

For the EDOS set, the most effective approach is the 5-shot classification, using the first five examples of each label from the training set with the Zephyr model, which achieves an M-F1 of 70.936% and a W-F1 of 77.514%. We can see that the 5-shot with a retrieval module based on Sentence Transformers has achieved an M-F1 of 70.544% and a W-F1 of 76.624%, which improves the ZSL performance, although overall it is inferior to the 5-shot with random examples. However, in detecting sexist texts (Sexism F1) it has the best result with 58.740%, beating the 5-shot by 0.578% and the ZSL by 4.111%.

In HatEval, the Zephyr model performs best with the 5-shot approach with an M-F1 of 70.115% and a W-F1 of 71.132%. However, for hate speech text, the StableBeluga-7b model performs best with the 5-shot approach with an F1 of 67.464%. In this case, the 5-shot approach with Sentence Transformers as the retrieval module does not improve the performance of the 5-shot, because the set selected by the retrieval module from the training set has a different structure and language patterns than the test set. Therefore, the examples inserted in the model prompt do not help to improve its performance compared to the 5-shot, but they did improve the M-F1 of the ZSL by 3.178%.

5.3 Comparison with Related Work

Based on the overview of the EDOS and HatEval tasks in binary classification of sexism and hate texts, the best result was selected to evaluate the reliability of our methods. Table 7 shows the best M-F1 of the task according to its ranking table and the best results of our approaches.

Table 7: Comparison table of the macro F1 score metric, showing the best result obtained with the ZSL and fine-tuning approach compared to the best macro F1 score according to the EDOS and HatEval task overview ranking table (Best M-F1)

Dataset	Best M-F1	Approach 1		Approach 2		
		Encoder-only	Encoder-decoder	Zero-shot	5-shot	5-shot ST
EDOS	87.460	83.913	**86.811**	66.304	**70.936**	70.544
HatEval	65.100	53.016	**57.453**	63.813	**70.115**	66.993

First, we observe that the 5-shot approach gave the best results in the null and few-shot classification using prompts in generative models. In this case, the best result was an M-F1 of 70.936 with a 5-shot approach using the Zephyr model, which is a decrease of 16.524% compared to the best EDOS result. However, in the fine-tuning approach, the Zephyr model, which is an encoder-decoder type, achieved the best result with 86.811%, only 0.649% lower. This result is very competitive and would place us in sixth place in its ranking.

Regarding the HatEval dataset, it is worth noting that the best result according to their ranking table was an M-F1 of 65.100. Our 5-shot approach has surpassed this result with a 70.115, which is an improvement of 5.015%. However, a similar situation occurs as discussed in the article [3], where the fine-tuning approach performs worse than a ZSL and FSL classification scenario. In our case, the best result of the fine-tuning approach was an M-F1 of 57.453.

Therefore, with these results, we can address RQ1, which asks whether LLM models can identify sexist and hateful texts online using ZSL and FSL approaches. The answer is that fine-tuned LLMs with an instructive dataset can directly identify sexist and hateful texts using prompts. In this case, the models perform better at identifying hateful text.

Tables 5 and 6 show that the best LLM for both ZSL, FSL, and fine-tuning is the Zephyr model, a fine-tuned version of Mistral-7B trained using Direct Preference Optimization (DPO) on a combination of synthetic and publicly available datasets. Thus, the answer to RQ2, which explores the best generative model, is Zephyr.

Finally, for RQ3, which asks whether fine-tuning a model for a classification task outperforms ZSL and FSL, the answer depends on the type of classification task. Looking at Table 7, it is clear

that ZSL and FSL outperform fine-tuning, similar to what was discussed in paper [3]. However, their performance is lower when identifying sexist texts, as shown by the EDOS results in Table 7.

6 Error Analysis

To better understand the behavior of the context-based learning and fine-tuning models, an error analysis is performed on the model that achieved the best macro F1 score over the test split of each dataset.

6.1 Sexism

For online sexism detection, the fine-tuning approach has achieved better results than those based on context learning. In this case, the Zephyr encoder-decoder model achieved the best result with 86.11%, beating the best result of context learning, which is a 5-shot learning of Zephyr, with 15.174%.

Table 8 shows a classification report of the Zephyr model in the two approaches. We can see that the fine-tuning based model predicts sexist speeches with an F1 of 79.97% and non-sexist speeches with an F1 of 93.64%, and from the recall values we can see that the model does not tend to predict a specific label. In contrast, the model based on context learning has a lower accuracy in detecting sexist texts, with an F1 of 58.16%. However, the prediction of non-sexist texts is significantly higher than that of sexist texts, with an F1 of 83.70%.

Table 8: Comparison of context learning and fine-tuning approaches using the classification report of the EDOS dataset. Both strategies use the Zephyr model

Label	Approach 1			Approach 2		
	Precision	Recall	F1	Precision	Recall	F1
Sexist	80.480	79.485	**79.979**	51.258	67.217	58.162
Not sexist	93.458	93.828	**93.643**	88.343	79.538	83.710
Macro avg.	86.969	86.657	**86.811**	69.801	73.377	70.936
Weighted avg.	90.311	90.350	**90.330**	79.350	76.550	77.514

Next, we analyze some of the false predictions made by the second approach (context learning) that the first approach (fine-tuning) is correct. In this sense, we want to analyze the weakness of the context learning approach. We found that the context learning approach tends to classify as sexist those documents that contain other types of hate-speech, such as racism. For example, the tweets "*Fuck the niggers and the Jews. Both have a history of eating white people and fuck you for suggesting I should get along with people who want to eat me.*" or "*#USA #Texas #Islam #Muslims #Islamization #ShariaLaw Sharia in Texas—Evil Muslim parents tried to murder their daughter with boiling oil after refusing arranged marriage.*"

We have also found that both the contextual learning approach and the fine-tuning approach tend to misidentify texts that contain figurative language or texts that are not inherently sexist. This is because sexism can often manifest itself subtly or implicitly in language without directly referring to gender or establishing specific roles for men or women. In these cases, it is important to consider the context and possible implications of the text in relation to gender norms and stereotypes. For example, the tweets "*10/10 with interior decorating skills like yours, girls will be falling all over you. I'm falling*

for you a little myself" or "*I feel the same way. It would be nice to have a wife and kids. But it would be horrible to have a vindictive ex and kids I never see. The risk is not worth the reward.*"

6.2 Hate Speech

In terms of hate speech detection, the Zephyr model has achieved the best results in both the context learning and fine-tuning approaches, with a macro F1 of 70.11% and 57.45%, respectively. In this case, the context learning based model has outperformed the fine-tuning based model, as shown in Table 9.

Table 9: Comparison of context learning and fine-tuning approaches using the classification report of the HatEval dataset. Both strategies use the Zephyr model

Label	Approach 1			Approach 2		
	Precision	Recall	F1	Precision	Recall	F1
Hate speech	50.825	95.397	**66.317**	68.330	59.762	63.760
Non hate speech	90.866	33.161	48.590	73.288	79.943	**76.471**
Macro avg.	70.845	64.279	57.453	70.809	69.852	**70.115**
Weighted avg.	74.049	59.300	56.035	71.206	71.467	**71.132**

Table 9 shows the classification report of the two models, and we can see that although Zephyr has achieved better overall performance with context learning, it has lower precision in identifying hate speech compared to the fine-tuning based model, with an F1 of 63.75 in context learning and 66.32% in fine-tuning. In addition, we can see that the main difference between the two approaches is in the detection of non-hate speech, in which case the FSL approach achieves an F1 of 76.47%, while fine-tuning achieves an F1 of 48.58%. This is because the standard model is already pre-trained with a large *corpus* and may contain hate-related text, so fine-tuning adds more noise to the model, which affects its performance in detecting non-hate speech, but improves the detection of hate speech.

Next, we analyze some of the wrong predictions made by the second approach (context learning) that the first approach (fine-tuning) is correct. In this sense, we want to analyze the weakness of the context learning approach. We found that the contextual learning approach tends to misclassify as non-hate speech those expressions against illegal immigration and the sustainability of immigration policies. However, HatEval identifies them as hate speech, so when the fine-tuning model has been trained on these examples, it is able to identify them as hate speech. For example, the tweets "*This is outrageous! #StopIllegalImmigration #MeritImmigration*" or "*Wake up #America! ! We are currently giving our #Country away!! We cannot sustain this and still try to #Help the #World... #Stopillegalimmigration #SundayMorning*."

We have also found that the fine-tuned model produces many false positives, taking into account those tweets that contain negative and derogatory content toward political or ideological groups, but do not rise to the level of hate speech. For example, the tweets "*@JudicialWatch My question is how do you resettle a refugee and why refugees go home when their country is at peace again illegal migrants don't.just because obummer allowed it doesn't mean it was legal*," "*GET this WORSE THAN SCUM OUT OF OUR COUNTRY! BUILD THAT WALL!!!! I cannot THANK ICE, ENOUGH, for ALL the GOOD they do, & HAVE DONE, for ALL AMERICANS—EVEN the ""Special"" ones who are TOO STUPID to KNOW IT. #LiberalismIsAMentalDisease #IStandWithICE #BuildThatWall #MAGA*

#2A," and "*@ICEgov The left can no longer be trusted. #LiberalismIsAMentalDisorder #GodBlessICE #MAGA #AmericaFirst #SendThemBack #BuildTheWall #DrainTheSwamp #walkaway @DNC*."

7 Conclusion and Further Work

LLMs are increasingly demonstrating their ability to understand natural language and solve complex tasks, especially through text generation. One of the relevant capabilities is in-context learning, i.e., the ability to receive instructions in natural language or task demonstrations to generate expected outputs for test instances without the need for additional training or gradient updates. This, LLMs have the capacity for ZSL and FSL learning.

Nowadays, with the exponential growth in the use of social media, a platform has been provided for the display of harmful and offensive online behavior, which is on the rise. As a result, new research and approaches have recently emerged to identify different types of offensive text. In this study, the capability of different LLM models, ranging from ZSL to fine-tuning approaches, has been investigated, and it has been observed that LLM models can identify sexist and hateful online texts using ZSL and FSL approaches through prompting. The encoder-decoder model called Zephyr performed best with the fine-tuning approach, scoring 86.811 on the EDOS test set and 57.453 on the HatEval test set. Finally, the models were found to perform well in detecting hate speech, surpassing the best score on the HatEval leaderboard.

A limitation of this research is that it relies on train validation data to obtain the best model. However, this approach may introduce bias in the model evaluation. To address this limitation, we propose to incorporate cross-validation. However, cross-validation requires more hardware and time resources that are not available.

As a future line of work, we intend to improve performance by using other models such as GPT-4, ChatGPT, and others, and by using alternative methods for FSL to find examples that are more similar to a particular class. In addition, improving detection performance by adding other text features such as emotion and sentiment is also a goal and performing a more detailed analysis of the performance of LLMs through RAG analysis.

Acknowledgement: None.

Funding Statement: This work is part of the research projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-A Way of Making Europe and LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. Mr. Ronghao Pan is supported by the Programa Investigo grant, funded by the Region of Murcia, the Spanish Ministry of Labour and Social Economy and the European Union-NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia (PRTR)."

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Rafael Valencia-García; data collection: Ronghao Pan; analysis and interpretation of results: Ronghao Pan, José Antonio García-Díaz; draft manuscript preparation: Ronghao Pan, José Antonio García-Díaz. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Public code repository and models deployed in Huggingface will be included in case of acceptance. EDOS dataset is available at https://github.com/rewire-online/edos. HatEval dataset is available at https://github.com/msang/hateval.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv preprint arXiv:2303.18223;2023. doi:10.48550/arXiv.2303.18223.
2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901. doi:10.48550/arXiv.2005.14165.
3. Plaza-del Arco FM, Nozza D, Hovy D. Leveraging label variation in large language models for zero-shot text classification. arXiv preprint arXiv:2307.12973; 2023. doi:10.48550/arXiv.2307.12973.
4. Su H, Kasai J, Wu CH, Shi W, Wang T, Xin J, et al. Selective annotation makes language models better few-shot learners. arXiv preprint arXiv:2209.01975; 2022. doi:10.48550/arXiv.2209.01975.
5. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682; 2022. doi:10.48550/arXiv.2206.07682.
6. Kirk H, Yin W, Vidgen B, Röttger P, Ojha K, Doğruöz AS, et al. SemEval-2023 task 10: explainable detection of online sexism. In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023); 2023; Toronto, Canada, Association for Computational Linguistics. p. 2193–210.
7. Basile V, Bosco C, Fersini E, Nozza D, Patti V, Rangel Pardo FM, et al. 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; 2019; Minneapolis, Minnesota, USA, Association for Computational Linguistics. p. 54–63.
8. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: open and efficient foundation language models. arXiv preprint arXiv:2302.13971; 2023. doi:10.48550/arXiv.2302.13971.
9. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020; p. 7871–80.
10. Workshop B, Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, et al. BLOOM: a 176b-parameter open-access multilingual language model. arXiv preprint arXiv: 2211.05100; 2022. doi:10.48550/arXiv.2211.05100.
11. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:6000–10.
13. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805; 2018. doi:10.48550/arXiv.1810.04805.
14. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692; 2019. doi:10.48550/arXiv.1907.11692.
15. Victor S, Albert W, Colin R, Stephen B, Lintang S, Zaid A, et al. Multitask prompted training enables zero-shot task generalization. In: International Conference on Learning Representations, 2022.
16. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824–37. doi:10.48550/arXiv.2201.11903.
17. Tang R, Han X, Jiang X, Hu X. Does synthetic data generation of llms help clinical text mining? arXiv preprint arXiv:2303.04360; 2023. doi:10.48550/arXiv.2303.04360.
18. Nov O, Singh N, Mann DM. Putting chatGPT's medical advice to the (Turing) test: Survey study. JMIR Med Educ. 2023;9:e46939. doi:10.2196/46939.
19. Yang K, Ji S, Zhang T, Xie Q, Ananiadou S. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. arXiv preprint arXiv:2304.03347; 2023.

20. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 2023;1:1–9. doi:10.1007/s00330-023-10213-1.
21. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. OpenAI; 2023.
22. Malinka K, Peresíni M, Firc A, Hujnák O, Janus F. On the educational impact of ChatGPT: is artificial intelligence ready to obtain a university degree? In: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1; 2023; New York, USA, Association for Computing Machinery. p. 47–53.
23. Blair-Stanek A, Holzenberger N, van Durme B. Can GPT-3 perform statutory reasoning? arXiv preprint arXiv:2302.06100; 2023. doi:10.48550/arXiv.2302.06100.
24. Trautmann D, Petrova A, Schilder F. Legal prompt engineering for multilingual legal judgement prediction. arXiv preprint arXiv:2212.02199; 2022. doi:10.48550/arXiv.2212.02199.
25. Choi JH, Hickman KE, Monahan A, Schwarcz D. ChatGPT goes to law school. J Legal Educ. 2022;387. doi:10.2139/ssrn.4335905.
26. Nay JJ. Law informs code: a legal informatics approach to aligning artificial intelligence with humans. Nw J Tech Intell Prop. 2022;20:3. doi:10.48550/arXiv.2209.13020.
27. Araci D. FinBERT: financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063; 2019. doi:10.48550/arXiv.1908.10063.
28. Salinas Alvarado JC, Verspoor K, Baldwin T. Domain adaption of named entity recognition to support credit risk assessment. In: Proceedings of the Australasian Language Technology Association Workshop. Parramatta, Australia; 2015. p. 84–90.
29. Wu S, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, et al. BloombergGPT: a large language model for finance. arXiv preprint arXiv:2303.17564; 2023. doi:10.48550/arXiv.2303.17564.
30. Chinivar S, Roopa MS, Arunalatha JS, Venugopal KR. Online offensive behaviour in socialmedia: detection approaches, comprehensive review and future directions. Entertain Comput. 2023;45:100544. doi:10.1016/j.entcom.2022.100544.
31. Beddiar DR, Jahan MS, Oussalah M. Data expansion using back translation and paraphrasing for hate speech detection. Online Soc Netw Media. 2021;24:100153. doi:10.1016/j.osnem.2021.100153.
32. Abburi H, Parikh P, Chhaya N, Varma V. Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. Data Sci Eng. 2021;6(4):359–79. doi:10.1007/s41019-021-00168-y.
33. Zhou X, Yong Y, Fan X, Ren G, Song Y, Diao Y, et al. Hate speech detection based on sentiment knowledge sharing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021; Association for Computational Linguistics. vol. 1, p. 7158–66.
34. Qureshi KA, Sabih M. Un-compromised credibility: social media based multi-class hate speech classification for text. IEEE Access. 2021;9:109465–77. doi:10.1109/ACCESS.2021.3101977.
35. Cheng L, Guo R, Silva YN, Hall D, Liu H. Modeling temporal patterns of cyberbullying detection with hierarchical attention networks. ACM/IMS Trans Data Sci. 2021;2(2):1–23. doi:10.1145/3441141.
36. Vijayaraghavan P, Larochelle H, Roy D. Interpretable multi-modal hate speech detection. arXiv preprint arXiv:2103.01616; 2021. doi:10.48550/arXiv.2103.01616.
37. Ge S, Cheng L, Liu H. Improving cyberbullying detection with user interaction. In: Proceedings of the Web Conference 2021; 2021. p. 496–506.
38. Khan MS, Malik MSI, Nadeem A. Detection of violence incitation expressions in urdu tweets using convolutional neural network. Expert Syst Appl. 2024;245:123174. doi:10.1016/j.eswa.2024.123174.
39. Bel-Enguix G, Gómez-Adorno H, Sierra G, Vásquez J, Andersen ST, Ojeda-Trueba S. Overview of HOMO-MEX at Iberlef 2023: hate speech detection in Online Messages directed towards the MEXican Spanish speaking LGBTQ$^+$ population. Procesamiento del Lenguaje Natural. 2023;71:361–70.

40. Rodríguez-Sánchez F, Carrillo-de -Albornoz J, Plaza L, Gonzalo J, Rosso P, Comet M, et al. Overview of exist 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural. 2021;67:195–207.
41. Plaza L, Carrillo-de -Albornoz J, Morante R, Amigó E, Gonzalo J, Spina D, et al. Overview of exist 2023: sexism identification in social networks. In: European Conference on Information Retrieval; 2023; Dublin: Springer; p. 593–9.
42. Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. arXiv preprint arXiv:2106.04554; 2021. doi:10.1016/j.aiopen.2022.10.001.
43. He P, Gao J, Chen W. DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543; 2021. doi:10.48550/arXiv.2111.09543.
44. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942; 2019. doi:10.48550/arXiv.1909.11942.
45. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108; 2019. doi:10.48550/arXiv.1910.01108.
46. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116; 2019. doi:10.48550/arXiv.1911.02116.
47. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. arXiv preprint arXiv:2310.06825; 2023. doi:10.48550/arXiv.2310.06825.
48. Tunstall L, Beeching E, Lambert N, Rajani N, Rasul K, Belkada Y, et al. Zephyr: direct distillation of LM alignment. arXiv preprint arXiv:2310.16944; 2023. doi:10.48550/arXiv.2310.16944.
49. Mukherjee S, Mitra A, Jawahar G, Agarwal S, Palangi H, Awadallah A. Orca: progressive learning from complex explanation traces of GPT-4. arXiv preprint arXiv:2306.02707; 2023. doi:10.48550/arXiv.2306.02707.
50. Ivison H, Wang Y, Pyatkin V, Lambert N, Peters M, Dasigi P, et al. Camels in a changing climate: enhancing lm adaptation with Tulu 2. arXiv preprint arXiv:2311.10702; 2023. doi:10.48550/arXiv.2311.10702.
51. Raza S, Ghuge S, Ding C, Pandya D. Fair enough: how can we develop and assess a fair-compliant dataset for large language models' training? arXiv preprint arXiv:2401.11033; 2024. doi:10.48550/arXiv.2401.11033.
52. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, et al. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652; 2021. doi:10.48550/arXiv.2109.01652.
53. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685; 2021. doi:10.48550/arXiv.2106.09685.