Multimedia Tools and Applications (2026) 85:323
https://doi.org/10.1007/s11042-026-21521-1

Clever domain adaptation strategies for BERT in the task of hostile-language detection

Emilio Villa-Cueva[1,2] · **Mario Ezra Aragón**[3] · **Adrián Pastor López-Monroy**[2] · **Fernando Sánchez-Vega**[2,4]

Received: 7 July 2025 / Revised: 21 January 2026 / Accepted: 23 March 2026

Abstract
Cyberbullying has experienced a surge in recent years, mainly due to the widespread adoption of social media platforms. This trend manifests in multiple ways, with hostile language being one of the most common. The latter underscores the urgent need for robust detection methods to address this issue effectively. To address this problem, we propose a novel pipeline to enhance hostile language detection in social media. Our approach consists of a combination of two ideas: First, we propose conducting a Domain Adaptation procedure to specialize the knowledge of a pre-trained BERT, making it more specialized in the domain of social media. For this adaptation, we modify the traditional random Masked Language Modeling technique and propose three novel strategies for selecting the subset of tokens to mask out cleverly. Second, we tailor an Adversarial Regularizer when fine-tuning the adapted BERT for specific hostile-language datasets. We evaluate the performance of our method for detecting hate speech, aggressiveness, offensiveness, and sexism. Our results show that the Domain Adaptation procedure significantly outperforms vanilla BERT, and the Adversarial Regularizer can lead to more robust fine-tuning, thereby enhancing performance. Moreover, we demonstrate that these methods can be used together to achieve an even more significant performance boost.

Keywords Hostile language · Domain adaptation · Social media · Text classification

1 Introduction

The past decade has witnessed an explosive proliferation of social media platforms, leading to an unprecedented surge in daily active users [1]. Social media has revolutionized human communication, fundamentally altering the way we interact with one another. These platforms, including Facebook, Twitter, Reddit, and numerous others, have become ubiquitous in modern society and were conceived to facilitate the dissemination of ideas, experiences, and opinions. Users of these platforms are constantly exposed to a diverse array of posts, messages, and replies, fostering an environment where interactions can sometimes become hostile. Such instances of hostility have the potential to inflict significant harm, especially

Extended author information available on the last page of the article

to individuals from vulnerable groups. For example, recent studies have shown a worrying correlation between cyberbullying and increased suicide rates, particularly among its victims, underscoring the urgent need for action [2]. In light of these findings, it becomes imperative for the scientific community to develop proactive systems capable of swiftly detecting and addressing such harmful behaviors before they escalate into devastating psychological or physical consequences. By leveraging technology and interdisciplinary collaboration, we can strive to create safer online spaces for all users, thereby mitigating the detrimental impact of cyberbullying on mental health and well-being.

In past years, models struggled to perform well when faced with data from domains different from their training data. Traditional approaches relied on manually crafted features or domain-specific resources, which were often labor-intensive and lacked scalability [3]. However, with the rise of deep learning, particularly transfer learning techniques such as fine-tuning pre-trained language models, domain adaptation in Natural Language Processing (NLP) has seen remarkable progress. Domain adaptation is a technique that enables a model trained on data from one domain (e.g., formal documents, such as books or news) to be adapted to perform well on data from a different domain (e.g., informal documents, such as social media posts). It helps NLP models specialize their understanding beyond the general context in which they were trained. Consider the case of pre-trained models, such as BERT [4], GPT [5], and their variants, which are trained on vast amounts of diverse text data, enabling them to capture rich linguistic patterns and semantic representations. By fine-tuning these pre-trained models on domain-specific data, researchers have achieved performance improvements across various domains without requiring extensive manual feature engineering. However, as this adaptation is usually done by adding a new layer at the end of the model and adjusting it to the task, sometimes important linguistic structures are lost.

Despite the success of pre-trained language models, their representations are primarily optimized for general-domain corpora (e.g., Wikipedia or books). When applied to hostile language detection on social media, these models face a significant domain mismatch caused by informal language, slang, obfuscated insults, and rapidly evolving vocabulary. Simple fine-tuning on small labeled datasets often fails to fully bridge this gap and may lead to unstable performance. In this context, rather than initially training the model, we propose some novel strategies to adjust the model's weights to align more closely with the language found in aggressive social media posts. Our approach involves conducting domain adaptation of the model before the fine-tuning phase for the classification task, similar to the one proposed in Han and Eisenstein [6]. This adaptation involves training the BERT model for the Masked Language Modeling task using data from both general-purpose (source domain) and social media-specific (target domain) sources in equal measure. For this adaptation, we explore various masking strategies, including selecting terms within a Term-Frequency interval and identifying the most essential tokens within the domain. Entropy selection of random and selected tokens for masking, measuring how hard it is to predict each token in the sentences, and specializing the model over the domain task. The main idea involves refining a pre-trained model using a relatively small corpus focused on the target domain [7]. Although this process incurs additional computational costs, it proves more economical than pre-training the model from scratch entirely. The main contributions of this work are summarized as follows:

- We propose a domain adaptation strategy that specializes a general-domain BERT mod-

el for social media language prior to task-specific fine-tuning, thereby reducing domain mismatch in hostile language detection.
- We introduce alternative masking strategies for Masked Language Modeling that leverage term frequency and prediction uncertainty to focus adaptation on domain-relevant tokens.
- We apply adversarial regularization during fine-tuning to improve robustness and reduce performance variance on small and noisy hostile-language datasets.
- We evaluate the proposed methods on multiple Spanish and bilingual benchmark datasets covering different forms of hostile language, demonstrating consistent performance improvements.

This document is structured as follows: In Section 2, we review previous approaches for text classification and hostile language detection. In Section 3, the concept of Domain Adaptation is reviewed on a deeper level, and we present our strategies for masking tokens. Section 4 explains the concept of an Adversarial Regularizer and how it is implemented in BERT. In Section 5.1, we present the bilingual and four Spanish datasets we use to evaluate our approach. In Section 6, we present the results of our approach, followed by a brief analysis of the results in Section 7. In Section 8, we address the ethical concerns and limitations of our approach. Finally, Section 9 includes the conclusions and future work.

2 Related work

The proliferation of hate speech on social media presents an escalating issue that perpetuates racial discrimination and erodes trust among individuals, fostering a climate conducive to physical violence and social division on a global scale. This problem generates an increasing effort to comprehend and anticipate this phenomenon, from analyzing behaviors [8] to domain generalization [9] and specialized lexicons [10]. Addressing this issue is challenging, primarily due to the deceptive nature of the problematic content and considerable noise within it. Researchers [11] have made strides in extracting relationships between textual data and disseminating hateful comments. However, further refinement and exploration are essential to deepen our understanding and develop more effective strategies for mitigating online hate speech.

Moreover, several research groups organized forums that addressed this topic by proposing shared tasks for identifying different types of hostile language. To name a few, TRAC-2 [12] in English, Dkhate [13] in Danish, and OSACT4 [14] in Arabic. Within the scope of this study, we focus our attention on Spanish datasets. Primarily motivated by the observation that taboo words may exhibit both hostile and non-hostile attributes, depending on the context in which they are employed.

The Spanish shared tasks have focused on different variations of hostile language: hate speech, offensiveness, aggressiveness, and sexism. HatEval [15] is a task designed to detect hate speech that targets a group based on inherent characteristics [16], specifically women and immigrants. The tweets in the HatEval dataset are in both Spain Spanish and Mexican Spanish. The Mex-A3T dataset [17] contains tweets primarily in Mexican Spanish and is designed to classify aggressive language, which exhibits anger and a readiness to harm others [18]. MeOffendEs [19] is another task intended to identify offensiveness that

causes upsetting feelings to others [20]. This task comprises two corpora, OffendMex and OffendEs, which are binary and multi-class datasets. OffendEs comprises posts in Spanish without filtering particular regions, whereas OffendMex has tweets in Mexican Spanish. Finally, the EXIST [21] task is designed to detect sexism in tweets in both English and Spanish, being the first task of its kind in this bilingual setting. Most participants proposed either traditional machine learning techniques, such as those presented in Pérez and Luque [22], which submitted the winning approach for HatEval, or Transformer models, including BERT [4]. For example, the works in Guzman-Silverio et al. [23] and Gómez-Espinosa et al. [24] obtained the best results in Mex-A3T and OffendMex, respectively, proposing solutions based on BETO [25], a BERT pre-trained in Spanish.

However, as several authors have pointed out, fine-tuning BERT[1] with a relatively small dataset can lead to high variance in performance. Sometimes, it even fails to converge, as it is susceptible to weight initialization and training data ordering [26, 27]. Additionally, BERT embeddings consider the context of each token [6]. This means that when the language domain of the classification task differs from the original pre-training domain, fine-tuning may not be optimal because a given word can have different meanings in different domains, depending on its context and usage.

Considering the latter issues, we aim to improve the performance of BETO classification for hostile language datasets by exploiting their specific social media language domain. To achieve this, we propose employing a Domain Adaptation strategy at the LMT level to adjust the model's weights, which were initially trained on a general domain corpus, to a new target social media domain before fine-tuning. We also propose addressing fine-tuning instability issues by exploiting an Adversarial Regularizer at the fine-tuning stage.

2.1 BERT in hostile language classification

Different models can be used to classify hostile language. Classic approaches typically involve extracting feature representations from the data and employing a classifier to perform the classification task. For example, Pérez and Luque [22] employed several representations (BoW, BoC, and Tweet Embeddings) in conjunction with an SVM, achieving the highest performance for the HatEval [15] task. Likewise, the best-performing team for classifying aggressiveness at Mex-A3T 2019 used character n-grams and word embeddings with an SVM and a multilayer perception. Nevertheless, BERT [4] models pre-trained in Spanish and then fine-tuned for the specific task have usually yielded the best results. BETO [25] –a BERT model pre-trained in Spanish– is used by the best-performing teams in Mex-A3T 2020[23], OffendES 2021 [24], and EXIST 2021 [21]. Anjum and Katarya [28] introduced a multilingual framework that combines BERT with a Multi-Layer Perceptron and a Profanity Check Technique. Their method involves code conversion, vector similarity analysis, and sentiment classification to accurately identify hate speech, even in poorly written or complex texts. Luu et al. [29] focused on Vietnamese hate speech detection by analyzing the impact of text pre-processing on BERTology models using two datasets: ViHSD and UIT-ViCTSD. They found that conventional pre-processing methods provide minimal improvements for transformer-based models. Dwivedy and Roy [30] proposed a multimodal hate speech detection architecture that fuses text and image features. Their model integrates transfer learning and LSTM techniques to classify social media posts.

[1] Even the base model.

Similarly, the research in Ghosh et al. [31] attempts to address the shortage of diversity in current hate speech detection datasets. The authors delve into a meticulously curated dataset of hate speech from Twitter, which encompasses various subjects, including terrorism, cyberbullying, natural disasters, and politics. They then test it using different models, including BERT, which yields the best results compared to base models. The dataset comprises 10,242 tweets categorized as either containing hate speech or not. BERT authors suggest fine-tuning it for a given task by plugging the task-specific inputs and outputs into the pre-trained model and then fine-tuning it in the task dataset end-to-end [4]. Despite its potential, a known issue with BERT is its sensitivity to weight initialization in the classification head and training data ordering, which leads to a high variance in performance at the moment of fine-tuning for specific tasks, especially if the training dataset is relatively small [27]. Due to the size of the hostile-language datasets, using a simple fine-tuned BERT can result in performance metrics with high variance. Some strategies for addressing this problem include leveraging an ensemble of BERT models with a voting scheme to compute the classification, as explored by Guzman-Silverio et al. [23]. Another approach is to employ Adversarial Regularization, as proposed by Jiang et al. [32]. In our study, we adopt the latter method during the fine-tuning phase to enhance the effectiveness of BERT in various hostile datasets.

Although this work focuses on text-based detection, related challenges have also been explored in other fields such as computer vision and multimodal learning. For instance, Wang et al. [33] introduces a multi-deliberation-based calibration strategy to address generalization issues in visual recognition, highlighting the importance of robustness under distribution shifts. Similarly, Susnjak et al. [34] investigates the domain-specific fine-tuning of large language models to enhance their adaptation to specialized domains. While these approaches target different modalities or objectives, they share a common motivation with our work: enhancing model adaptation and robustness when transitioning to new domains.

3 Domain Adaptation (DA) for a social-media language domain

The term language domain (or textual domain) describes a language distribution that characterizes a particular field or topic [35]. Thus, a specific vocabulary is commonly employed, and its interpretation might have specific conventions. Therefore, a given phrase may have different meanings when contextualized in different textual domains. Because of this, training Language Models in a language domain and evaluating them in another may decrease their performance. BERT is pre-trained for a general-purpose language domain[2] that performs satisfactorily for several tasks. In this case, its contextualized embeddings are adjusted to the general domain and not necessarily the task's specific language domain [6].

Nonetheless, previous research has shown that pre-training BERT in a language domain closely aligned with the target task's domain can yield substantial performance enhancements [36]. However, undertaking BERT's pre-training from scratch is computationally demanding, and sufficient data in the desired target domain may be limited. Popularly, BERT models have three forms of fine-tuning. The Masked Language Model (MLM), next sentence prediction, and adding a layer in the CLS token. MLM consists of masking a portion of the input tokens in a sentence at random and then asking the model to predict the

[2] We refer to the general-purpose language domain as that of Wikipedia, books, and web crawling.

masked tokens. Next sentence prediction, on the other hand, involves the model receiving two sentences and predicting whether the sentences are related and whether the input sentence is the next one. Lastly, add a layer after the CLS token to learn the weights of the new task.

This can be alleviated through Domain Adaptation strategies, such as AdaptaBERT [6], which has demonstrated that a pre-trained general model can be adapted to perform better on a target domain using a smaller amount of target domain data. This is achieved by fine-tuning the model using MLM with data from both the target and source domains. The masked examples are created following the process that is used for pre-training BERT: 15% of the input tokens are randomly masked, and the model is trained for three epochs by minimizing the negative log probability of the masked sample:

$$\min_{S \in D} - \sum \log \Pr(S \mid S^{msk}) \tag{1}$$

Where S is the unmasked sentence and S^{msk} is the randomly masked input.

Regarding the data, the examples are randomly drawn from both the source and target domains.

3.1 Proposed approach

To adjust the already pre-trained model weights in a general domain to a textual domain more closely aligned with hostile social media posts, we propose conducting domain adaptation on the model before the fine-tuning stage for the classification task. Although this DA procedure incurs some additional computational cost, it remains more cost-effective than pre-training the model from scratch.

For this, we follow an approach similar to AdaptaBERT [6]: we train the BERT model on the Masked Language Modelling task using source domain data (general-purpose) and target domain data (social media) in equal proportions. The intuitive idea behind this approach is to adapt the existing weights in a clever manner. Figure 1 illustrates the main steps of our proposed strategy, which we further expand upon in the following. For the DA procedure in Spanish, we use data from the following sources:

- The CC-100 dataset [37] for the source domain data.

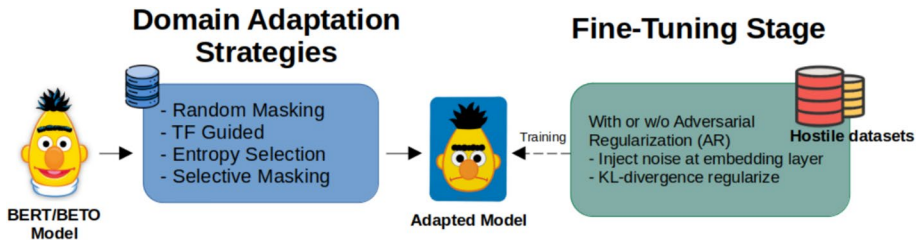

Fig. 1 Domain adaptation and fine-tuning pipeline. A BERT/BETO model is adapted using masking-based strategies, followed by fine-tuning on hostile datasets

- The *TwitterSentimentDataset*[3] for social media.

Since we also evaluate this method on a bilingual dataset (EXIST), another Domain Adaptation uses a different bilingual corpus. In this case, we use the same corpus for Spanish tweets described above and include the following training data in English:

- The *Wiki-40B* [38] dataset for source domain.
- The *Sentiment140* [39] dataset for social-domain data.

For each of the four datasets, we randomly selected approximately 250,000 sentences, resulting in a corpus of around 500,000 samples for the Spanish DA and 1 million samples for the bilingual DA.

While random masking is effortless to implement, it remains a naive approach for unsupervised learning. Masking out randomly chosen tokens may result in selecting examples that do not significantly contribute to the transition from the language domain to the desired target domain. Ideally, we would like to choose masks that challenge the language model as much as possible. To achieve this, we explored additional masking strategies that could selectively mask a group of tokens, thereby transitioning the model's domain as much as possible in each training iteration. This resulted in a better-performing and more robust adapted model.

Using the unlabeled corpus mentioned above, we evaluate AdaptaBERT Random Masking and the Uncertainty Based Strategy proposed by Vu et al. [40] as DA baselines. Then, we propose three strategies for mask selection: Term-Frequency Guided, Entropy Selection of Random Samples, and Selective Masking.

3.1.1 Random Masking (RND)

It follows the AdaptaBERT Masking Scheme, which involves further pre-training BERT using a procedure similar to the one employed during BERT's original training. For each sentence fed into BERT, 15% of the tokens are randomly replaced by the mask token.

3.1.2 Uncertainty-Based Strategy. (UBS)

This approach, proposed by Vu et al. [40], consists of feeding BERT the unmasked sentence S and then computing the entropy in each of the output tokens, given by:

$$H\left(t_i \mid S\right) = -\sum_{j=0}^{|V|} \Pr\left(t_{i,j} \mid S\right) \log\left(\Pr\left(t_{i,j} \mid S\right)\right) \tag{2}$$

Where $\Pr\left(t_{i,j} \mid S\right)$ is the probability output of BERT of the j-th token of the vocabulary at the i-th token of the sentence.

We then select 15% of tokens with the highest entropy values as the masks for the UDA. This approach aims to mask out the tokens for which BERT is more *unconfident* about. However, this does not consider combinations of masked tokens that may increase entropy.

[3] Available in this repository: https://github.com/garnachod/TwitterSentimentDataset

3.1.3 Term-Frequency Guided (TF)

We propose a DA strategy that randomly selects tokens within a Term-Frequency interval. Firstly, we compute the TF values for each unique token in the entire Domain Adaptation Dataset. This step is done before carrying out the DA procedure. We can retrieve intervals from arbitrarily defined quantiles from the computed TF values. During training, the masks are drawn randomly from tokens with TF values within an interval specified by a pair of quantiles.

The main difference between this procedure and Random Masking is that we "discard" tokens whose TF value is not within the given interval. The intuition of this approach is that the most important tokens for a given domain may fall within a given TF quantile. Thus, we experiment with different intervals.

3.1.4 Entropy Selection of Random Samples (ES-RS)

For this method, we create K masking configurations[4] S_k^{msk} for each sentence S, feed them to BERT and compute the output entropies $H\left(t_i \mid S_k^{msk}\right)$ of the masked tokens using equation 2. Then, we select the $|msk|$ tokens with the highest entropy values out of all the K masking configurations, where $|msk|$ is the number of masked tokens that correspond to 15% of the sentence length. We will use this subset of tokens as our masks for that specific sentence in the MLM task.

3.1.5 Selective masking

Finally, we propose a method that selects tokens that are especially hard to predict by BERT: For each token in each sentence, we create an input with that single token masked and feed it to BERT. We then measure its output entropy $H\left(t_i \mid S_k^{msk}\right)$ and evaluate if it was correctly or incorrectly predicted. Then, we generate our sentence with $|msk|$ masks, following a 1:3 proportion of correctly predicted to incorrectly predicted tokens, selecting those with the higher entropy values.

As one might expect, this masking strategy requires significantly more computational time than the others, since every token in every sentence must be masked and predicted.

3.2 Methodology, motivation, and linguistic hypothesis

The novelty of our approach lies in rethinking how domain adaptation is performed for hostile language detection. We introduce linguistically motivated masking strategies that guide the Masked Language Modeling objective toward domain-relevant tokens, and we systematically combine this adaptation with adversarial regularization during fine-tuning. To the best of our knowledge, this combination has not been previously explored for hostile language detection in social media.

To bridge the gap between general-domain pretraining and the specific linguistic characteristics of hostile language on social media, we explore Domain Adaptation (DA) using Masked Language Modeling (MLM). While traditional DA approaches rely on random masking, this naive strategy may mask tokens that contribute little to domain-specific

[4] Sets of randomly selected tokens (15% of the sentence) to mask out.

knowledge transfer. Our goal is to develop more informed and linguistically motivated masking strategies that can better guide the model to adapt to the target domain. This motivation is grounded in a key linguistic hypothesis: *"Frequent terms and uncertain tokens are likely to be domain-relevant features in social media text."*

Frequent terms as domain anchors Social media discourse–especially in hostile contexts–frequently reuses certain words, such as slang, informal variants, insults, and group-specific jargon. These terms often appear in mid-to-high frequency bands within domain-specific corpora and are less common or differently contextualized in general-language data. We propose that frequent tokens capture the statistical and thematic salience of the domain, serving as lexical anchors for adaptation. This insight motivates our Term-Frequency Guided (TF) strategy, which masks tokens selected from a specified TF interval. By focusing MLM on these domain-relevant tokens, we encourage the model to adjust its internal representations better to reflect the domain's vocabulary and usage patterns.

Uncertainty as a signal of domain mismatch In contrast, some tokens, though not necessarily frequent, may be difficult for a general-domain model to predict, indicating a mismatch between the model's learned knowledge and the target domain. These uncertain tokens often include polysemous words used in novel ways, coded or obfuscated hate speech, or non-standard syntactic forms. This motivates a family of uncertainty-driven masking strategies:

- Uncertainty-Based Strategy (UBS): Selects the top 15% of tokens in a sentence with the highest output entropy when unmasked, focusing on the most ambiguous terms.
- Entropy Selection of Random Samples (ES-RS): Creates multiple random masking configurations per sentence, computes entropies for each, and selects the configuration that yields the highest uncertainty across masked tokens. This allows us to choose the most informative token combinations dynamically.
- Selective Masking: For each token, we individually mask it and compute its prediction entropy. We then select a mix of correctly and incorrectly predicted tokens with high entropy values. This approach identifies the most challenging parts of a sentence, pushing the model to improve on specific weaknesses.

Balancing simplicity and efficiency Together, these masking strategies form a spectrum from simple (Random Masking) to computationally intensive (Selective Masking). Each strategy reflects a trade-off between implementation cost and adaptation quality. While Random Masking is efficient, it lacks guidance. In contrast, strategies like ES-RS and Selective Masking more deliberately target domain-specific linguistic signals at the cost of increased computation.

By aligning our methodology with these linguistic hypotheses, we aim to maximize the utility of unlabeled domain data, helping the model learn more robust, domain-aware representations–ultimately improving its ability to detect hostile language in noisy, user-generated content.

4 Adversarial regularization

Due to the high capacity of BERT, overfitting may occur on specific tasks, making it crucial to train a model that correctly generalizes unseen data. Regularizers can improve a model's generalization without affecting its training error [41]. Adversarial training achieves this by training a regularizer that harnesses adversarial examples [42]. This effectively trains a more robust model by injecting slight noise oriented in the "worst-case" direction at the model inputs, making the model more resistant to small changes in the inputs. The general form of the loss function with a regularizer is expressed below.

$$J\left(\mathrm{x}, y, \theta\right) = L\left(\mathrm{x}, y, \theta\right) + \alpha R_{\mathrm{adv}}\left(\mathrm{x}, \theta\right) \tag{3}$$

Where $L\left(\cdot\right)$ is the Cross-Entropy Loss to optimize during training and $R_{\mathrm{adv}}\left(\cdot\right)$ is the regularizer, in our case, an adversarial regularizer.

Several authors have implemented adversarial regularizers, obtaining promising results [32, 42–44]. For our task, we require a regularizer for language models; we chose the method proposed in Jiang et al. [32], which injects the directed noise at the embedding layer of the model, since it was suitable for our application. This method regularizes classification tasks by using a symmetrized Kullback–Leibler Divergence (which, in this case, measures the difference between the output distributions of the model when given a noisy input and the original one) to find the direction of the noise that maximizes this divergence.

This directed noise is contained inside a hypersphere of radius ϵ, following the expression in equations 4 and 5.

$$R_{\mathrm{adv}}\left(\mathrm{x}, \theta\right) = \max_{\|\widetilde{x} - x\| < \epsilon} L_D\left(\mathrm{B}\left(\mathrm{x}, \theta\right), \mathrm{B}\left(\widetilde{\mathbf{x}}, \theta\right)\right) \tag{4}$$

$$\widetilde{\mathbf{x}} = \mathbf{x_r} + \eta \mathrm{sign}\left(\nabla_{\mathbf{x_r}} L_D\left(\mathrm{B}\left(\mathbf{x_r}\right), \mathrm{B}\left(\mathrm{x}\right)\right)\right) \tag{5}$$

Where $\mathrm{B}(\cdot)$ is the output of BERT, $L_D(x_1, x_2) = D_{KL}(x_1, x_2) + D_{KL}(x_2, x_1)$ is the symmetrized KL divergence and $\mathbf{x_r}$ is a noised input such that $L_D(\cdot)$ derivative is not zero, the noised input follows the form $\mathbf{x_r} = \mathrm{x} + \mathrm{r}$ where $\mathrm{r} \sim \mathcal{N}(0, 1e - 0.5)$. This noise can't be injected at the model's input because it is a tokenized sentence, so it is introduced after the embedding layer.

As noted by Jiang et al. [32], adversarial training can benefit the model in classification tasks by *smoothing out* the decision boundaries between classes, which leads to improved performance when classifying new examples. Hence, we utilize this procedure for hostile-language classification tasks.

5 Experimental settings

5.1 Hostile language datasets

Since our work focuses on enhancing the performance of BERT in hostile language detection tasks, we evaluate our strategies using five datasets. Four are in Spanish, and one is in

a bilingual setting (Spanish-English). These have been labeled for hate speech, aggressive language, offensive language, and sexist content. Table 1 presents a general overview of these datasets.

5.1.1 OffendEs and OffendMex

The MeOffendEs task [19], presented at IberLEF 2021, aims to detect *offensive language* in Spanish social media and is divided into the OffendEs and OffendMex corpora. The first contains comments from various social media platforms, primarily in general Spanish. It is labeled in a multi-class setting according to the following categories: (OFP) Offensive, where the target is a person; (OFG) Offensive, where the target is a group; (OFO) Offensive, where the target is neither a person nor a group; and (NO) Non-Offensive. This corpus comprises more than 30,000 posts, with a subset labeled by three annotators and another subset labeled by ten annotators using a majority vote scheme.

The OffendMex corpus focuses on tweets in Mexican Spanish, with binary labels indicating whether a tweet is offensive or non-offensive. It has been manually labeled and contains more than 7,000 tweets.

5.1.2 Mex-A3T

Introduced at IberLEF 2020, Mex-A3T [17] presented the task of detecting *aggressive language* in Mexican Spanish. For this, the authors compiled a corpus of more than 10,000 tweets, filtered to exclude rude words and controversial hashtags related to sexism, homophobia, and discrimination, resulting in a dataset containing several tweets labeled as either aggressive or non-aggressive. The authors mention that labeling was challenging because tweets needed to be interpreted in their particular context to determine whether they were aggressive or used inappropriate vocabulary correctly.

Table 1 List of hostile language datasets

Dataset	URL	Type
OffendEs/OffendMex	competitions.codalab.org/competitions/28679	hostile and aggressiveness
MEX-A3T	sites.google.com/view/mex-a3t/home	aggressiveness
Hateval	competitions.codalab.org/competitions/19935	xenophobia and sexism
EXIST	nlp.uned.es/exist2021/	sexism

5.1.3 Hateval

HatEval [15] is one of the most popular tasks presented at SemEval-2019. This dataset contains *Hate Speech* mainly aimed at women and immigrants, and includes 13,000 English and 6,000 Spanish tweets. The authors mention that the tweets were collected by monitoring potential victims of hate accounts, downloading the history of identified haters, and filtering tweets by specific keywords. The data was then labeled into the following three categories: (1) Hate Speech (HS), a binary value that indicates if the tweet contains HS or not, if it does (2) Target Range, which indicates whether the target is a generic group or an individual and (3) Aggressiveness, another binary value that indicates if the tweet containing HS is aggressive or not. The task was divided into two subtasks: Subtask A, which involved a binary classification problem for the HS label only, and Subtask B, which entailed classifying two other classes, Target Range and Aggressiveness.

In this work, we focus on Subtask A in Spanish, and from here onward, we will refer to this specific Subtask when referring to the HatEval.

5.1.4 EXIST

EXIST [21] is a shared task introduced at IberLEF 2021 that aims to *detect sexism* in a broad sense. Its corpus is formed by Spanish and English posts collected from Twitter and *Gab.com* and labeled under the supervision of experts in gender issues. The tasks were composed of two subtasks: subtask one consisted of binary classification (sexist or non-sexist), and subtask two consisted of multi-class classification, in which the sexist label was divided into *Ideological and inequality, Stereotyping and dominance, Objectification, Sexual violence,* and *Misogyny and non-sexual violence*, thus it was multi-class classification problem with six classes. For subtask one, the tweet (or gab) was labeled as *sexist* if it expressed sexist behavior or a sexist discourse; otherwise, it was classified as non-sexist.

The EXIST corpus comprises more than 11,300 posts in Spanish and English, in a nearly equal proportion. In this work, we evaluate subtask 1 (binary classification); from here onward, any mention of EXIST will refer to EXIST 2021 subtask one.

5.2 Model configuration

5.2.1 Pre-processing

The preprocessing of tweets involves converting all text to lowercase, removing emojis, and replacing user mentions and URLs with the placeholders '@user' and '', respectively.

5.2.2 Implementation settings

For Domain Adaptation, we train with a batch size of 128 and a learning rate of 5.0×10^{-5} for four epochs. For the fine-tuning stage, we train for between 3 and 4 epochs, optimizing with a learning rate in the range of 1×10^{-5} to 2×10^{-5}. All models were trained using one NVIDIA TITAN RTX GPU. It is important to notice that, given modern hardware, BERT-based models can operate efficiently at inference time, even on CPUs. In high-throughput environments, batch processing enables practical near real-time deployment.

The hyperparameters for the adversarial regularization were selected based on performance on the development sets (see Table 3). The adversarial regularization term is implemented following the approach introduced by Jiang et al. [32]. We refer the reader to this work for a detailed discussion of optimization and stability considerations.

5.2.3 Development partitions

We randomly sample 10% of the training data of all the datasets described above to use as a *development partition* and conduct experiments on the effectiveness of the proposed DA and AR procedures.

6 Evaluation results

6.1 Effect of domain adaptation and regularization parameter on development partitions

We utilize the development partitions described in Section 5.2.3 to evaluate the effectiveness of the proposed strategies and fine-tune their hyperparameters. Considering that most DA strategies contain a random factor that may yield different results for a single method (aside from the random ordering of the training data), we train four Domain Adaptations for each DA strategy. We fine-tune and evaluate five runs of them in our in-house partition to determine their mean performance and standard deviation, reported in Table 2.

We assess the performance of Random, UBS, TF-Guided, and ES-RS on Spanish tasks using the same set of initialization seeds for the classification layer. Performance results of the Selective DA are reported in Section 6.2, as only one DA was carried out using this strategy.

As we can observe in Fig. 2, performance in HatEval was much more challenging to improve through Domain Adaptation than OffendMex and Mex-A3T. This may be due to HatEval containing many racial slurs and xenophobic terms, which are less likely to be present in the corpora used for domain adaptation.

For TF-Guided DA, we found that larger TF values performed better, and tokens with TF values in the quantile range of $[Q(0.55), Q(1.0)]$ resulted in highly effective masks. Note that masking the most common terms enhances the domain adaptation process because these terms are likely widely used across social media. This process is also helpful in establishing the context for social media jargon using stop words.

Table 2 Performance of different DA methods in the development partitions

Macro F1 - development set	Mex-A3T		HatEval		OffendMex	
Domain Adaptation	Mean	Std	Mean	Std	Mean	Std
No DA	84.227	0.647	85.632	0.706	77.585	0.651
Random	85.621 (+1.654%)	0.752	86.188(+0.650%)	0.593	79.995(+3.106%)	0.855
UBS	85.795 (+1.861%)	0.802	85.656(+0.027%)	0.667	79.653(+2.665%)	0.858
ES-RS (k=5,top k)	87.045 (+3.345%)	0.701	86.005(+0.435%)	0.523	80.219(+3.394%)	0.874
TF Q(0.55,1.0)	86.491(+2.688%)	0.650	86.152(+0.607%)	0.725	79.781(+2.830%)	0.890

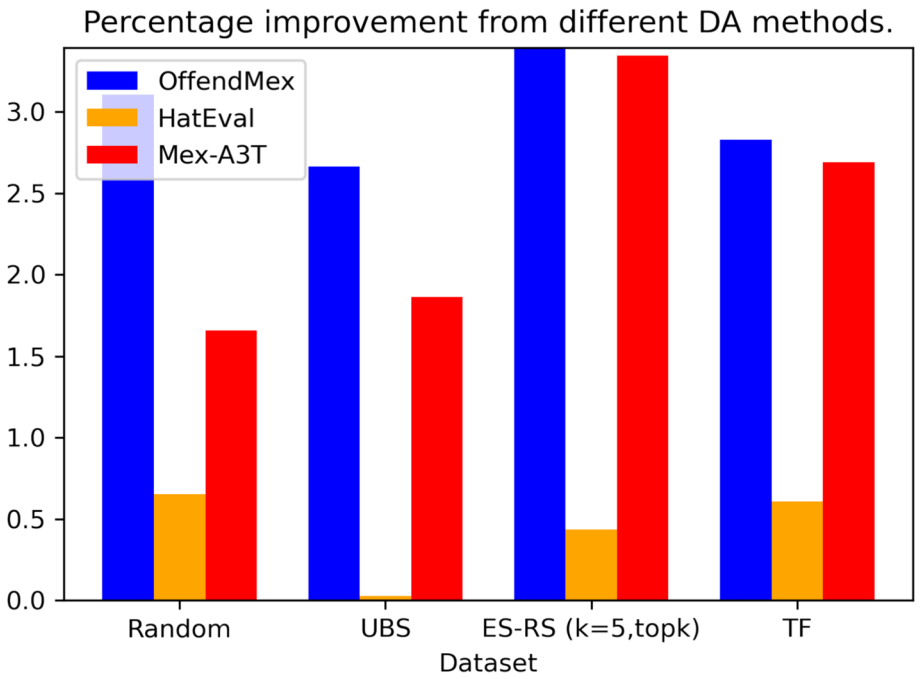

Fig. 2 Performance of different Domain Adaptation Strategies in the development set

Table 3 Effects of performance of the regularization parameter λ on the development partitions using a general-domain BERT

α	OffendMex		Mex-A3T		HatEval	
	Mean	Std	Mean	Std	Mean	Std
0	77.524	0.378	84.648	0.802	85.451	0.49
0.5	78.314	1.259	-	-	86.049	0.407
1	78.397	0.83	85.628	0.304	85.855	0.485

As for ES-RS, the best performance results were obtained using $k = 5$; we observed that using a larger value, such as $k = 10$, decreased performance. Both of these strategies yielded relatively stable performance improvements across all datasets. Both of these strategies achieved relatively stable performance improvements over the random baseline across all datasets, suggesting that, despite their differences, each approach effectively focuses resources on words relevant to adapting to the target domain.

We also evaluate the effect of different values of α for the regularizer by evaluating several non-adapted models in the development partitions (See Table 3). We observe that different datasets require different values of α. Therefore, we select the best-performing α for each dataset in the development partition to use in the test partitions.

6.2 Performance results in test partitions

We fine-tuned five models for each configuration[5] and evaluated them in the *official* test partitions of the datasets, the mean macro-F1 scores are reported in Table 4 for the Spanish (we include the best reported results for each task) and in Table 5 for the bilingual task. Results show that the best performance was achieved by combining Domain Adaptation and Adversarial Regularization.

Our Domain Adaptations based on alternative masking strategies improved over random masking in most cases. However, contrary to the development set in which ES-RS yielded the highest average improvement, TF-Guided DA improved performance the most in the test set of OffendMex and HatEval. At the same time, ES-RS achieved the highest improvement in OffendEs and EXIST. Finally, Selective Masking also yielded good results but only outperformed the other methods in the Mex-A3T dataset.

We found that α was the hyperparameter that most significantly affected performance for the Adversarial Regularizer. In general, we found that the best value depended on the dataset. At the same time, $\alpha = 1.0$ was a suitable value for OffendES, OffendMex, and Mex-A3T, while $\alpha = 0.5$ yielded the best results for HatEval, and $\alpha = 1.5$ was the optimal value for EXIST. We generally found $\eta = 0.001$ and $\epsilon = 5.0 \times 10^{-5}$ to be suitable hyperparameters for all datasets.

Table 4 F1 scores of our different models in the testing partitions

	OffendEs		OffendMex		Mex-A3T		HatEval	
Model	Macro F1 - mean	std	Macro F1 - mean	std	Macro F1 - mean	std	Macro F1 - mean	std
Vanilla BERT	71.905	0.869	69.14	0.538	84.95	0.223	76.453	0.22
Best Reported	73.240 [19]	-	70.260 [19]	-	85.960 [17]	-	73.000 [15]	-
BERT +DA$_{ES-RS}$	76.277 (+6.08%)	0.423	71.012 (+2.707%)	0.564	85.818 (+1.022%)	0.235	77.017 (+0.739%)	0.923
BERT +DA$_{TF}$	75.893 (+5.55%)	0.407	73.473 (+6.266%)	0.206	86.344 (+1.641%)	0.252	77.394 (+1.231%)	1.159
BERT +DA$_{Random}$	75.705 (+5.28%)	0.335	69.879 (+1.07%)	0.27	86.54 (+1.87%)	0.243	77.217 (+1.00%)	0.827
BERT +DA$_{Selective}$	75.987 (+5.68%)	0.142	71.061 (+2.78%)	0.428	86.108 (+1.36%)	0.238	76.879 (+0.56%)	0.36
Model +AR								
Vanilla BERT	73.958 (+2.86%)	0.303	69.629 (+0.706%)	0.266	84.7 (-0.294%)	0.266	76.385 (-0.089%)	0.915
BERT +DA$_{ES-RS}$	**77.465 (+7.73%)**	0.494	71.542(+3.474%)	0.184	86.67 (+2.025%)	0.268	77.341 (+1.162%)	0.466
BERT +DA$_{TF}$	76.912 (+6.96%)	0.16	**73.559 (+6.39%)**	0.37	86.866 (+2.255%)	0.174	**77.905 (+1.899%)**	0.984
BERT +DA$_{Random}$	76.846 (+6.87%)	0.458	70.752 (+2.33%)	0.369	86.63 (+1.98%)	0.21	77.789 (+1.75%)	0.44
BERT +DA$_{Selective}$	77.124 (+7.26%)	0.349	71.469 (+3.37%)	0.664	**87.17 (+2.61%)**	0.298	77.629 (+1.54%)	0.65
RoBERTuito	77.893	0.217	72.667	0.407	87.602	0.162	80.128	0.515

[5] *Configuration* refers to a specific DA strategy with or without AR.

Table 5 Macro F1 scores of different strategies in EXIST partition

EXIST (5-runs mean)	F1(mean)	F1(std)
Vanilla mBERT	74.885	0.376
mBERT + AR	75.764 (+1.17%)	0.301
mBERT +DA(ES-RS)	76.604 (+2.3%)	0.39
mBERT +DA (ES-RS) + AR	77.328 (+3.26%)	0.227
mBERT +DA (TF)	76.681 (+2.4%)	0.173
mBERT +DA (TF) + AR	76.711 (+2.44%)	0.280
mBERT +DA(Random)	76.375 (+1.99%)	0.421
mBERT +DA (Random) + AR	76.753 (+2.49%)	0.188
mBERT +DA (Selective)	76.104 (+1.63%)	0.187
mBERT +DA (Selective) + AR	76.971 (+2.79%)	0.261
BERT (es)	75.396	0.471
BERT (es) + AR	75.998 (+0.8%)	0.274
BERT (es) + DA(ES-RS)	77.313 (+2.54%)	0.197
BERT (es) + DA(ES-RS) + AR	**77.751 (+3.12%)**	0.11
BERT (es) + DA (TF)	77.213 (+2.41%)	0.424
BERT (es) + DA (TF) + AR	77.368 (+2.62%)	0.116
BERT (es) + DA(Random)	77.288 (+2.51%)	0.379
BERT (es) + DA(Random) + AR	77.328 (+2.56%)	0.27

We found that the evaluated Domain Adaptation strategies complement the Adversarial Regularizer. Therefore, using both can potentially increase the performance of vanilla BERT models by a significant margin. Specifically, we observed that both ES-RS and TF-Guided approaches, combined with an Adversarial Regularizer, yielded the best results overall, suggesting that there is still room for improvement in a traditional BERT model without the need for pre-training domain-specific architectures.

In Table 4, we also include the evaluations of the RoBERTuito model [45], which utilizes the RoBERTa [46] architecture and has been trained from scratch using Spanish tweets. Therefore, it represents an upper bound in terms of performance increase due to the language domain for our models. Although RoBERTuito outperforms our models by some degree in most cases, it is essential to note that we did not pre-trained BERT, so our approach used around 0.04% of target-domain samples compared to them[6] and most of our Domain Adaptations took around 6 hours in a Titan RTX GPU[7], compared to the three weeks in a GCP v3-8 TPU that RoBERTuito authors needed to train their models.

7 Results analysis

We briefly analyzed our results in the testing partitions to suggest a few key ideas of why the Domain Adaptation procedure improves performance more in some datasets than others. The experimental results support the contributions of this work. Domain-adapted models consistently outperform vanilla BERT across various datasets, while informed masking strategies generally outperform random masking in most cases. Additionally, adversarial regularization complements domain adaptation by enhancing robustness and improving per-

[6] Our DA dataset contained 250k Spanish tweets compared to theirs, which consisted on 622M tweets

[7] Except Selective Masking which took around two days

formance stability. We focused our analysis on the four Spanish datasets, as the bilingual EXIST uses a different Domain Adaptation Approach.

7.1 Complete token representation

To analyze the complexity of sentences in different corpora, we examine the number of words that have a complete token representation, meaning they are not split into subword tokens. This analysis indicates the sentence complexity within each corpus. For instance, the model may represent a general-domain sentence with five words using six tokens. In contrast, a social media domain sentence of the same length may require eight tokens, thus increasing the modeling complexity of that sentence.

Since we based our adapted models on BETO, we use the same tokenizer [25]. In Fig. 3, we can observe that most datasets have a complete token representation in around 87% of their words, except OffendEs, which has a higher value. This may be because it contains posts from different social media platforms, not just Twitter.

Additionally, we observe that the Domain Adaptation corpora provide a more comprehensive representation of the source domain (general-domain Spanish) compared to the target domain (tweets). Also, the task datasets on which we evaluated our models have fewer split words than the Domain Adaptation target corpus, suggesting that the domain may not be as closely aligned as we initially thought.

The reduced quantity of complete word tokens within the target domain's corpora suggests that, despite our efforts to effectively shift the domain from the source domain to the target domain by updating the model weights, an inherent drawback persists from the tokenization process.

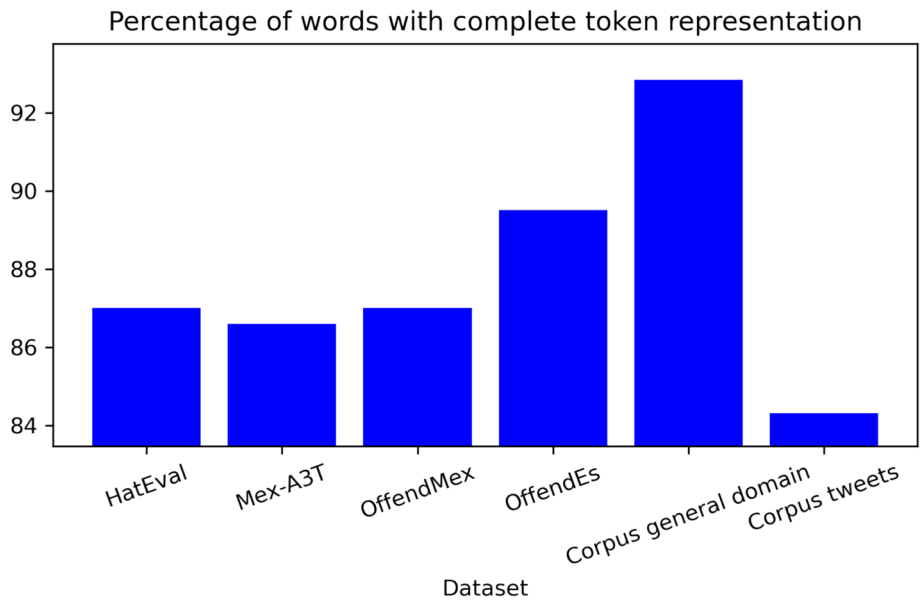

Fig. 3 Percentage of words in the dataset with complete token representation

7.2 Shared words with DA corpus

To evaluate the similarity between the task datasets and the Domain Adaptation corpora, for each task dataset, we computed the percentage of all the tokens in the task dataset that are present in the Domain Adaptation corpora (source domain, target domain, and joint domains) in the following way:

$$\#_{i=0}^{|V|}\left(\#\left(t_i \in C_{\text{tsk}} \mid t_i \in C_{\text{DA}}\right)\right) \times \frac{100}{\mid C_{\text{tsk}} \mid} \tag{6}$$

Where C_{tsk} is the task dataset, C_{DA} the Domain Adapatation corpus and $|t_i|$ is the total number of occurrences of token t_i in C_{tsk}. The obtained percentages are exhibited in Fig. 4. Additionally, we include the average performance improvement due to Domain Adaptation in Fig. 5 for comparison purposes.

We noticed that the words shared with the tweets (target domain) corpus behave very similarly to the performance increase. Datasets with a higher percentage of shared words tend to benefit from a greater performance gain. Furthermore, for the HatEval dataset — which turns out to be the dataset with the smallest increase in performance— there is a clear difference between the test set and the training set shared words, hinting that the test set could be slightly less similar to the Domain Adaptation target domain than the training set. These outcomes suggest that a higher percentage of shared words with the target domain corpus may be a crucial factor for successful Domain Adaptation of BERT.

7.3 Error analysis through attention visualization

To better understand the remaining errors of our models, we analyze which tokens receive higher attention in representative misclassified and correctly classified examples. In Figs. 6 and 7, we can observe the mean values of the last transformer block of the BERT model when classifying some selected examples. This allows us to shed light on the attention mechanism within the model and identify which parts of the tweets are most important for the model's prediction.

The models adapted for tweets show a broader distribution of attention per token, aligning with expectations for the concise nature of tweets, where contextually significant information tends to be more condensed than in longer texts. For the task of aggressiveness detection in Mexican Spanish (Fig. 6), notable attention is given to specific expressions like *putos* (faggots) and *al chile* (seriously), reflecting their use in Mexican Spanish social media posts. Interestingly, certain curse words adapted for social media slang, such as *seenga tu madre* (a wordplay substituting a vulgar insult with seen), are effectively interpreted in their intended context by the models.

In analyzing the examples from the OffendEs dataset (Fig. 7), which focuses on identifying offensiveness in Mexican Spanish, specifically targeting groups rather than individuals, the adapted models demonstrate increased attention to terms like *gentuza* (a slang term for despicable people). Additionally, expressions carrying obscene connotations, including *mover el orto* (shake her ass) and *tenia mejor papo* (she had the best pussy), are recognized with notably higher attention scores, underscoring the models' capacity to grasp the significance of such slang within its context.

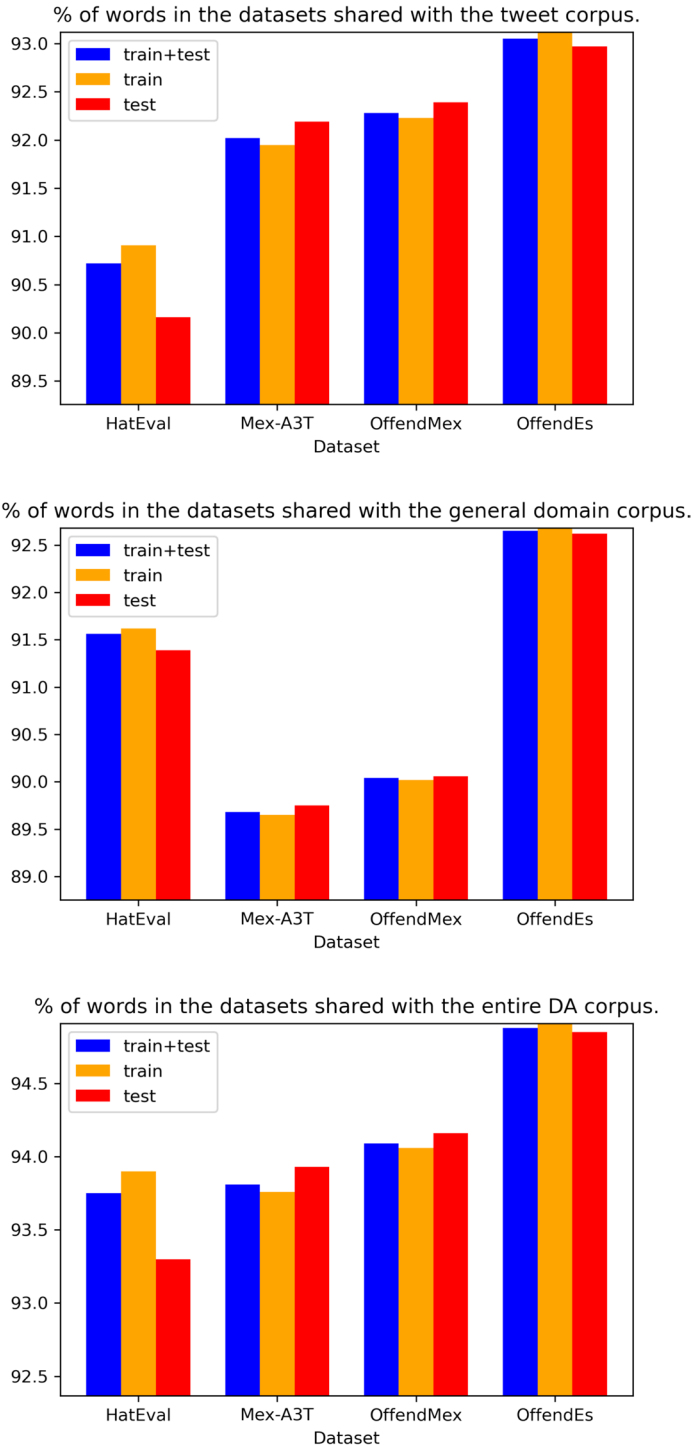

Fig. 4 Percentage of words in the evaluation datasets shared with the domain adaptation corpora

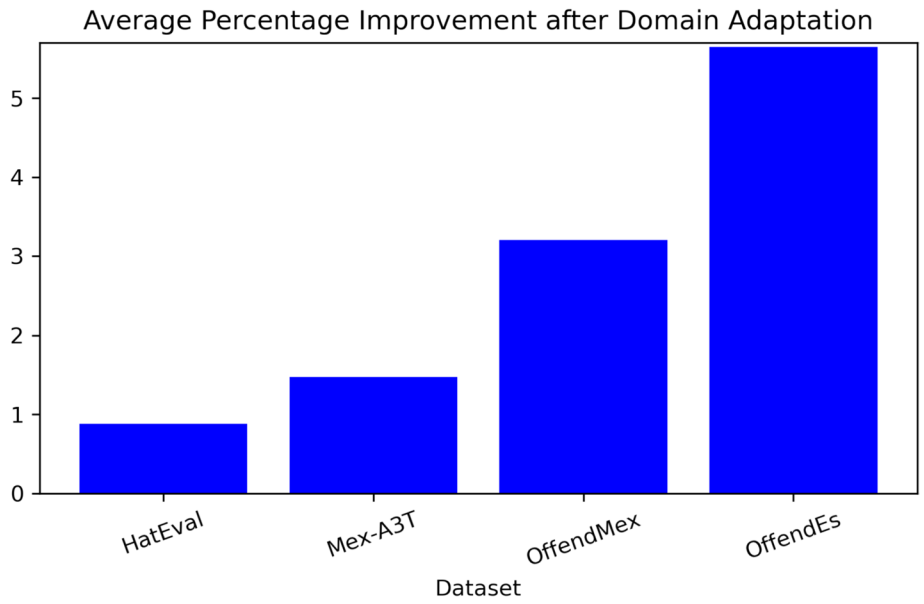

Fig. 5 Percentage improvement per dataset

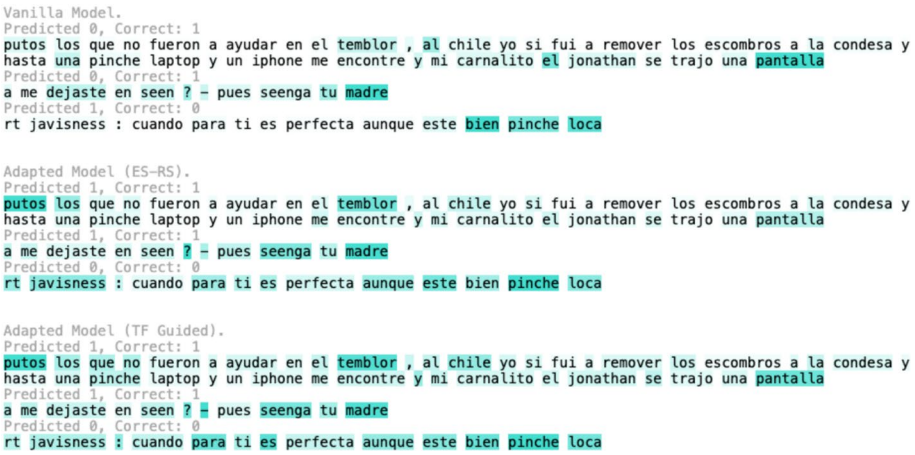

Fig. 6 Attention scores at different words on the MexA3T dataset. In this figure, we compare the model with TF-guided and ES-RS adaptations against the vanilla model in a few selected instances

7.4 Analyzing the pseudo-perplexities (PPPL) of adapted models

BERT's bidirectional architecture computes probabilities of masked tokens using its previous and future tokens, which makes calculating its perplexity impossible because the probability of a given token is not conditioned only by its previous tokens, as in other Language Models.

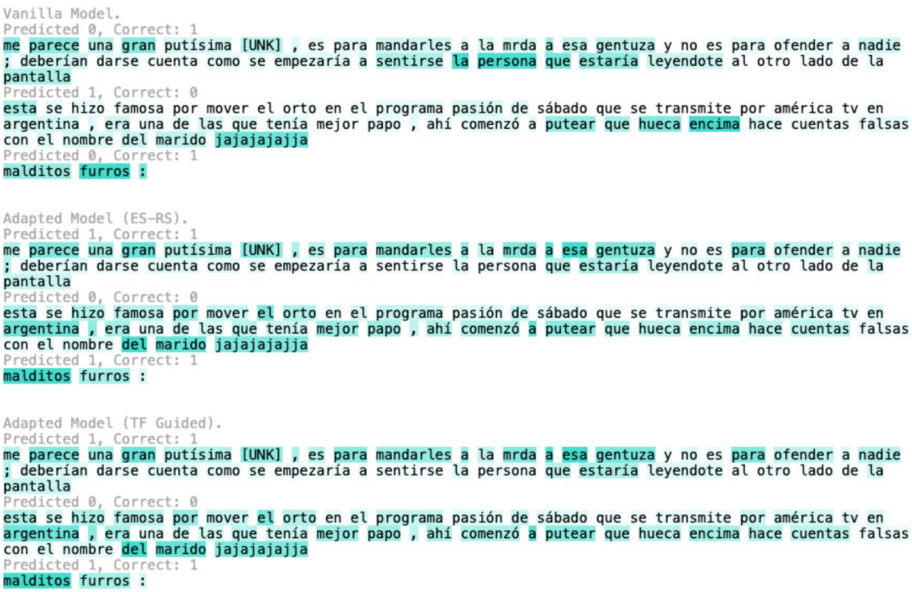

Fig. 7 Attention scores at different words on the OffendEs dataset. In this figure, we compare the model with TF-guided and ES-RS adaptations against the vanilla model in a few selected instances

Table 6 Pseudo-perplexities of BERT after the DA stage

	Mex-A3T	OffendMex	OffendEs	Hateval
Vanilla BERT	166.151	167.087	72.903	248.85
BERT +DA$_{Random}$	12.627	12.046	13.694	9.815
BERT +DA$_{TF}$	17.975	17.219	20.782	14.833
BERT +DA$_{ES-RS}$	12.449	11.832	13.442	9.406
BERT +DA$_{Selective}$	12.378	11.782	13.839	9.341

Some approaches estimate this metric considering the bidirectional nature of BERT, such as the *pseudo-perplexity (PPPL)* introduced in Salazar et al. [47]. In summary, it estimates token probabilities in a sentence by masking each token individually and then predicting it based on its previous and future context. We examined how the DA procedure alters this metric in the corpora of our evaluation tasks, where a reduced PPPL value in social media corpora would indicate that our models have successfully transitioned to a social media domain.

After the DA phase, a substantial drop in PPPL can be observed in the corpora (As shown in Tables 6 and 7 for Spanish and bilingual DA, respectively). Compared to other methods, TF-Guided DA resulted in a smaller reduction of PPPL, possibly due to masking only tokens with higher TF values during the DA (See Section 3). This did not affect the model's performance when fine-tuning it for the classification tasks.

Table 7 Pseudo-perplexities of BERT after the DA stage on the bilingual dataset

	EXIST
mBERT Vanilla	21.56819956
mBERT +DA$_{Random}$	8.466563468
mBERT +DA$_{TF}$	14.66534382
mBERT +DA$_{ES-RS}$	8.372404715
mBERT +DA$_{Selective}$	8.990085487

8 Ethical statement and limitations

It is crucial to acknowledge that human contributors have curated and annotated the datasets utilized as benchmarks in this study. Their categorization of content as hateful, offensive, aggressive, or sexist is inherently subjective and susceptible to cultural and geographical biases. These biases can significantly influence the interpretation and labeling of data points. Even in cases where multiple annotators have contributed to labeling the datasets, their perspectives may be influenced by shared cultural backgrounds, especially if they originate from similar geographical regions. Consequently, their labeling criteria may exhibit consistency shaped by their social contexts.

Furthermore, automated hostile language detection systems may reflect societal biases present in the data and may disproportionately affect certain demographic groups. These risks highlight the importance of transparency, continuous evaluation, and human oversight when deploying such models in content moderation settings. For instance, within the HatEval dataset, a notable proportion of the identified hateful tweets often target immigrants from specific regions, such as South America or North Africa. Consequently, models trained on this dataset may demonstrate heightened proficiency in detecting hateful language directed towards these communities compared to identifying hate speech targeting other demographics. As a result, the model inadvertently absorbs any biases inherent in the dataset during the training process. Therefore, researchers and practitioners must exercise caution when employing such datasets and be mindful of the potential biases that may be embedded within them. Moreover, efforts should be made to diversify dataset annotations by incorporating perspectives from individuals representing a broad spectrum of cultures and backgrounds, thereby mitigating the impact of inherent biases on model performance and generalization.

Another limitation of this study is the focus on a selected subset of hostile language types, specifically offensive language, aggressive language, sexist language, and xenophobic language. While these categories represent a significant portion of online hostility and allow for meaningful analysis across varied forms of abuse, they do not encompass all possible types of hostile language. For instance, racial discrimination and other nuanced or emerging forms of hostility were not explicitly addressed. Future work could expand the scope to include a broader range of hostile language types, thereby validating the findings further and generalizing them. Nonetheless, the current study offers valuable insights by testing across multiple distinct hostility categories, capturing a broad spectrum of online aggression.

All experiments were conducted on NVIDIA Titan RTX GPUs (24GB memory). The computational requirements of the proposed model are comparable to standard BERT-based classifiers. While optimization techniques, such as distillation or model compression, could

further reduce resource consumption, exploring these strategies is beyond the scope of this work.

From an environmental standpoint, while Domain Adaptation entails greater computational demands than mere fine-tuning, transitioning a pre-trained language model to a new domain consumes less energy than training it from scratch for a particular language. Remarkably, this approach yields significant performance enhancements compared to general-domain pre-trained models. Furthermore, research indicates that the initial training of the BERT base model can incur costs ranging from $\$3,751$ to $\$12,571$ USD when utilizing Cloud Compute Services, accompanied by an emission of approximately $650kg$ of CO_2 into the atmosphere [48]. Domain Adaptation is a compelling solution for languages lacking a pre-trained model tailored to a specific domain. Such languages may require additional resources for the comprehensive pre-training of expansive models tailored to domain-specific tasks.

We acknowledge the complexity inherent in analyzing content sourced from social media platforms. This data's very nature raises concerns about privacy and ethical propriety. It is crucial to emphasize that our research methodology exclusively utilized pre-existing, publicly accessible datasets. We abstained from any direct engagement or interaction with users across social media platforms, thus ensuring the preservation of user privacy and integrity. The datasets utilized in our study were obtained from the official sites, a reputable source known for its commitment to ethical data acquisition practices. Furthermore, we meticulously safeguarded the anonymity and confidentiality of all dataset participants. Our research endeavors were conducted with the utmost respect for ethical principles, ensuring transparency, integrity, and compliance with relevant regulatory frameworks. By upholding these standards, we aimed to mitigate potential ethical concerns while advancing our understanding of the phenomena under investigation.

9 Conclusions and future work

This work presents a practical domain adaptation approach that incorporates social media information into the model, along with new strategies for the masked language modeling process, which further improve the results. The findings demonstrate a consistent enhancement in performance across hostile language datasets when employing Data Augmentation (DA). Notably, we discern that the method of selecting masked tokens for the Masked Language Modeling Task has a significant impact on the outcome. Specifically, our investigation reveals that tokens chosen through the *TF* and *ES-RS* strategies result in the most substantial improvements. We think this is because these strategies successfully capture and incorporate important linguistic patterns related to the target domain. Furthermore, integrating Adversarial Regularization (AR) with DA proves to be efficacious, showcasing performance strides across various adaptations and models. However, the efficacy of this approach is contingent upon the strength of regularization, highlighting the importance of meticulous hyperparameter selection.

In summary, this study highlights the effectiveness of combining DA and AR within BERT to improve hostile language detection on social media platforms. These synergistic techniques enhance model robustness and highlight the nuanced interplay between data augmentation and regularization in improving model performance.

For future work, we aim to evaluate the effectiveness of the proposed domain adaptation strategies across other transformer-based architectures, particularly encoder-decoder models such as T5 [49] and BART [50]. In these models, the masking strategies presented here could be adapted to their denoising objective, potentially improving their performance in specific domains. Additionally, we propose grounding the mask selection process in a set of terms specific to social media or aggression, identified through a linguistic perspective. This approach could further refine the domain adaptation process by concentrating resources on the most relevant terms for the target task. Although our experiments are limited to Twitter (now X), widening the evaluation to other platforms (e.g., Reddit, Facebook, or forum-based communities) also remains an important direction for future research, as platform-specific conventions and linguistic variation may affect model performance. Lastly, we plan to explore the use of adapters or other parameter-efficient techniques to reduce the memory requirements of the method, making it more accessible for resource-constrained environments. Furthermore, we will extend the scope of our analysis to encompass other critical categories, thereby creating a more inclusive and generalizable framework for detecting hostile language. Finally, we plan to address the challenge of imbalanced data distribution, which can disproportionately affect the model's performance across different classes. Exploring advanced sampling techniques, cost-sensitive learning, and augmentation strategies tailored to minority classes will be key steps in mitigating this issue. Addressing class imbalance effectively contributes to a fairer and more robust classification system, particularly in real-world applications where skewed distributions are prevalent.

Acknowledgements We thank CONAHCYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies and CIMAT Bajio Supercomputing Laboratory (#300832). Villa-Cueva (CVU 1019520) thanks CONAHCYT for the support through the master's degree scholarship at CIMAT. This work was supported by the project CBF-2025-I-4384, "Grandes Modelos de Lenguaje Especializados para Detectar Ciberacoso y Violencia Digital", approved under the Ciencia Básica y de Frontera 2025 call of SECIHTI, Mexico. Mario Ezra Aragón thanks the support obtained from MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22, supported by ERDF), Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19, supported by ERDF), and the support obtained from the Juan de la Cierva Grant (JDC2023-052296-I), funded by MCIN/AEI/10.13039/501100011033 and by the FSE+. Sanchez-Vega acknowledges CONAHCYT / SECIHTI's support through the program "Investigadoras e Investigadores por México" (Project ID 11989, No. 1311).

Author Contributions **Emilio Villa-Cueva:** Conceptualization, Formal analysis, Methodology, Writing – original draft. **Mario Ezra Aragón:** Validation, Writing – review and editing. **Fernando Sánchez-Vega:** Formal analysis, Supervision, Writing – review and editing. **Adrián Pastor López-Monroy:** Conceptualization, Formal analysis, Supervision, Funding acquisition, Project administration, Writing – review and editing.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was funded by Xunta de Galicia, Ministerio de Ciencia e Innovación (Spain), and CONAHCYT (Mexico).

Data Availability Data available on request from the authors.

Declarations

Compliance with ethical standards This manuscript is the author's original work and has not been published or submitted simultaneously elsewhere. All authors have reviewed the manuscript and agree to its submission.

Conflict of interest On behalf of all authors, the corresponding author declares that there is no conflict of interest.

Competing interests The authors have no competing interests relevant to the content of this article.

Ethical Approval This study did not involve human or animal subjects; therefore, ethical approval and informed consent were not applicable. We only used public data and did not contact social media users.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

References

1. Ortiz-Ospina E (2019) The rise of social media. Our world in data. https://ourworldindata.org/rise-of-social-media
2. Luxton DD, June JD, Fairall JM (2012) Social media and suicide: A public health perspective. Am J Public Health 102(S2):195–200. https://doi.org/10.2105/ajph.2011.300608
3. Zhao Z, Alzubaidi L, Zhang J, Duan Y, Gu Y (2024) A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. Expert Syst Appl 242:122807. https://doi.org/10.1016/j.eswa.2023.122807
4. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423
5. Radford A, Narasimhan K (2018) Improving language understanding by generative pre-training. https://api.semanticscholar.org/CorpusID:49313245
6. Han X, Eisenstein J (2019) Unsupervised domain adaptation of contextualized embeddings for sequence labeling
7. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA (2020) Don't stop pretraining: Adapt language models to domains and tasks. In: Jurafsky D, Chai J, Schluter N, Tetreault J (eds) Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Online, pp 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740
8. Noorian Z, Ghenai A, Moradisani H, Zarrinkalam F, Alavijeh SZ (2024) User-centric modeling of online hate through the lens of psycholinguistic patterns and behaviors in social media. IEEE Trans Comput Soc Syst 1–13. https://doi.org/10.1109/TCSS.2024.3359010
9. Awal MR, Lee RK-W, Tanwar E, Garg T, Chakraborty T (2024) Model-agnostic meta-learning for multilingual hate speech detection. IEEE Trans Comput Soc Syst 11(1):1086–1095. https://doi.org/10.1109/TCSS.2023.3252401
10. Ahmed U, Lin JC-W (2022) Deep explainable hate speech active learning on social-media data. IEEE Trans Comput Soc Syst 1–11. https://doi.org/10.1109/TCSS.2022.3165136
11. Maity K, Sen T, Saha S, Bhattacharyya P (2024) Mtbullygnn: A graph neural network-based multitask framework for cyberbullying detection. IEEE Trans Comput Soc Syst 11(1):849–858. https://doi.org/10.1109/TCSS.2022.3230974

12. Bhattacharya S, Singh S, Kumar R, Bansal A, Bhagat A, Dawer Y, Lahiri B, Ojha AK (2020) Developing a multilingual annotated corpus of misogyny and aggression. In: Proceedings of the second workshop on trolling, aggression and cyberbullying. European Language Resources Association (ELRA), Marseille, France, pp 158–168. https://www.aclweb.org/anthology/2020.trac2-1.25
13. Sigurbergsson GI, Derczynski L (2020) Offensive language and hate speech detection for Danish. In: Proceedings of the twelfth language resources and evaluation conference. European Language Resources Association, Marseille, France, pp 3498–3508. https://aclanthology.org/2020.lrec-1.430
14. Mubarak H, Darwish K, Magdy W, Elsayed T, Al-Khalifa H (2020) Overview of OSACT4 Arabic offensive language detection shared task. In: Proceedings of the 4th workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection. European Language Resource Association, Marseille, France, pp 48–52. https://aclanthology.org/2020.osact-1.7
15. Basile V, Bosco C, Fersini E, Nozza D, Patti V, Pardo FMR, Rosso P, Sanguinetti M (2019) 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. Association for Computational Linguistics. https://doi.org/10.18653/v1/s19-2007
16. Nations U: What is hate speech? | United Nations — un.org. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech. Accessed 05 May 2023
17. Aragón ME, Jarquín-Vásquez HJ, Montes-y-Gómez M, Escalante HJ, Pineda LV, Gómez-Adorno H, Posadas-Durán JPF, Bel-Enguix G (2020) Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In: IberLEF@SEPLN
18. Aggressive. Houghton mifflin harcourt publishing company. https://dictionary.cambridge.org/dictionary/english/aggressive Accessed 04 May 2023
19. Plaza-del-Arco FM, Casavantes M, Escalante HJ, Martín Valdivia MT, Montejo Ráez A, Gómez M, Jarquín-Vásquez H, Pineda L (2021) Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants
20. Offensive. https://dictionary.cambridge.org/dictionary/english/aggressive Accessed 04 May 2023
21. Rodríguez-Sánchez F, Carrillo-de-Albornoz J, Plaza L, Gonzalo J, Rosso P, Comet M, Donoso T (2021) Overview of EXIST 2021: sEXism Identification in Social neTworks. Procesamiento del Lenguaje Natural 67:195–207
22. Pérez JM, Luque FM (2019) Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification. In: Proceedings of the 13th international workshop on semantic evaluation. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp 64–69. https://doi.org/10.18653/v1/S19-2008
23. Guzman-Silverio M, Balderas-Paredes Á, López-Monroy AP (2020) Transformers and data augmentation for aggressiveness detection in mexican Spanish. In: IberLEF@SEPLN
24. Gómez-Espinosa V, niz-Sánchez VM, López-Monroy AP (2021) Transformers pipeline for offensiveness detection in mexican Spanish social media. In: IberLEF@SEPLN
25. Cañete J, Chaperon G, Fuentes R, Ho J-H, Kang H, Pérez J (2020) Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020
26. Mosbach M, Andriushchenko M, Klakow D (2021) On the Stability of Fine-tuning BERT: misconceptions, explanations, and strong baselines
27. Dodge J, Ilharco G, Schwartz R, Farhadi A, Hajishirzi H, Smith N (2020) Fine-tuning pretrained language models: weight initializations, data orders, and early stopping
28. Anjum, Katarya R (2024) Hatedetector: Multilingual technique for the analysis and detection of online hate speech in social networks. Multimed Tools Appl 83(16):48021–48048. https://doi.org/10.1007/s11042-023-16598-x
29. Luu ST, Van Nguyen K, Nguyen NL-T (2024) An approach of data augmentation to improve the performance of bertology models for vietnamese hate speech detection. Multimed Tools Appl 83(19):56763–56783. https://doi.org/10.1007/s11042-023-16968-5
30. Dwivedy V, Roy PK (2023) Deep feature fusion for hate speech detection: a transfer learning approach. Multimed Tools Appl 82(23):36279–36301. https://doi.org/10.1007/s11042-023-14850-y
31. Ghosh S, Ekbal A, Bhattacharyya P, Saha T, Kumar A, Srivastava S (2023) Sehc: A benchmark setup to identify online hate speech in English. IEEE Trans Comput Soc Syst 10(2):760–770. https://doi.org/10.1109/TCSS.2022.3157474
32. Jiang H, He P, Chen W, Liu X, Gao J, Zhao T (2020) SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.197
33. Wang Y, Li J, Zhang H, Zhang J, Wan F, Qiu A, Gao Z (2025) Fedmdd: Multi-deliberation based calibration for federated long-tailed learning. Knowl-Based Syst 323:113741. https://doi.org/10.1016/j.knosys.2025.113741

34. Susnjak T, Hwang P, Reyes N, Barczak ALC, McIntosh T, Ranathunga S (2025) Automating research synthesis with domain-specific large language model fine-tuning. ACM Trans Knowl Discov Data 19(3). https://doi.org/10.1145/3715964
35. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA (2020) Don't stop pretraining: adapt language models to domains and tasks
36. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2019) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz682
37. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2020) Unsupervised cross-lingual representation learning at scale
38. Guo M, Dai Z, Vrandecic D, Al-Rfou R (2020) Wiki-40b: Multilingual language model dataset. In: LREC 2020. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.296.pdf
39. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. Processing 150
40. Vu T-T, Phung D, Haffari G (2020) Effective unsupervised domain adaptation with adversarially trained language models
41. Goodfellow IJ, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge, MA, USA. http://www.deeplearningbook.org
42. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples
43. Miyato T, Maeda S-i, Koyama M, Ishii S (2017) Virtual adversarial training: A regularization method for supervised and semi-supervised learning. https://doi.org/10.48550/ARXIV.1704.03976.
44. Zhang H, Yu Y, Jiao J, Xing EP, Ghaoui LE, Jordan MI (2019) Theoretically principled trade-off between robustness and accuracy. https://doi.org/10.48550/ARXIV.1901.08573
45. Pérez JM, Furman DA, Alemany LA, Luque F (2021) Robertuito: a pre-trained language model for social media text in Spanish. arXiv:2111.09453
46. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A robustly optimized BERT pretraining approach. https://doi.org/10.48550/ARXIV.1907.11692
47. Salazar J, Liang D, Nguyen TQ, Kirchhoff K (2019) Pseudolikelihood reranking with masked language models. arXiv:1910.14659
48. Strubell E, Ganesh A, McCallum A (2019) Energy and policy considerations for deep learning in NLP. https://doi.org/10.48550/ARXIV.1906.02243
49. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(140):1–67
50. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2019) Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461

Authors and Affiliations

Emilio Villa-Cueva[1,2] · Mario Ezra Aragón[3] · Adrián Pastor López-Monroy[2] · Fernando Sánchez-Vega[2,4]

✉ Mario Ezra Aragón
ezra.aragon@usc.es

Emilio Villa-Cueva
emilio.villa@mbzuai.ac.ae

Adrián Pastor López-Monroy
pastor.lopez@cimat.mx

Fernando Sánchez-Vega
fernando.sanchez@cimat.mx

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

[2] Department of Computer Science, Mathematics Research Center (CIMAT), A.C, Guanajuato, Mexico

[3] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela (USC), Santiago de Compostela, Spain

[4] Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico City, Mexico