Nuanced Toxicity Detection in Spanish: A New Corpus and Benchmark Study

Alba María Mármol-Romero[1], **Robiert Sepúlveda-Torres**[2], **Estela Saquete**[2], **María-Teresa Martín-Valdivia**[1], **Alfonso Ureña**[1]
[1]Computer Science Department, SINAI, CEATIC, University of Jaén, Spain
[2]Department of Software and Computing Systems, University of Alicante, Spain
[1]{amarmol, maite, laurena}@ujaen.es
[2]{rsepulveda, stela}@dlsi.ua.es

Abstract

The rise of toxic content on digital platforms has intensified the demand for automatic moderation tools. While English has benefited from large-scale annotated corpora, Spanish remains under-resourced, particularly for nuanced cases of toxicity such as irony, sarcasm, or indirect aggression. We present an extended version of the NECOS-TOX corpus, SocialTox, comprising 4,011 Spanish comments collected from 16 major news outlets. Each comment is annotated across three levels of toxicity (Non-Toxic, Slightly Toxic, and Toxic), following an iterative annotation protocol that achieved substantial inter-annotator agreement (k = 0.74). To reduce annotation costs while maintaining quality, we employed a human-in-the-loop active learning strategy, with manual correction of model pre-labels. We benchmarked the dataset with traditional machine learning (ML) methods, domain-specific transformers, and instruction-tuned large language models (LLMs). Results show that compact encoder models (e.g., RoBERTa-base-bne, 125M parameters) perform on par with much larger models (e.g., LLaMA-3.1-8B), underscoring the value of in-domain adaptation over raw scale. Our error analysis highlights persistent challenges in distinguishing subtle forms of toxicity, especially sarcasm and implicit insults, and reveals entity-related biases that motivate anonymization strategies. The dataset and trained models are released publicly.

1 Introduction

The exponential growth of user-generated content on digital platforms has intensified the urgency of developing reliable systems for detecting and mitigating toxic language online. Toxic comments—defined here as offensive, aggressive, or harmful expressions—can disrupt constructive dialogue, polarize communities, and undermine digital well-being. While substantial progress has been achieved for English through the availability of large annotated datasets and deep learning methods, many other widely spoken languages remain underrepresented. Spanish, despite being the second most spoken native language worldwide, still lacks high-quality and nuanced corpora for toxicity detection. Models trained on English data often fail to transfer effectively, leaving Spanish-speaking communities without tools adapted to their linguistic and cultural realities.

Existing Spanish corpora address this gap only partially. For example, DETOXIS (Taulé et al., 2021) focuses on immigration-related toxicity, while NECOS-TOX (López-Úbeda et al., 2024) links toxicity with discourse constructiveness but is restricted to a single outlet. These resources are valuable but limited in terms of representativeness and in their ability to capture subtle, borderline cases such as irony, sarcasm, or delegitimization, phenomena that are pervasive in online political and social discourse yet remain poorly operationalized in Natural Language Processing (NLP).

In this paper, we present an expanded and diversified version of NECOS-TOX, called SocialTox, that directly targets these challenges. The corpus contains 4,011 Spanish comments drawn from 16 major news outlets, annotated following an iterative, human-in-the-loop protocol that combines model-assisted pre-labeling with expert adjudication[1]. Our refined annotation scheme distinguishes three levels of toxicity—*Non-Toxic*, *Slightly Toxic*, and *Toxic*—designed to capture not only explicit insults but also the more elusive forms of rhetorical aggression. Table 1 illustrates representative examples. This corpus emphasizes diversity and nuance, aiming to expose linguistic subtleties that current models consistently struggle with.

[1]**Disclaimer.** The toxic comments reproduced in this work are drawn from publicly available data and are included exclusively for the purpose of scientific analysis. They do not reflect the views of the authors, their institutions, or the conference organizers. Readers may find some examples offensive.

Findings of the Association for Computational Linguistics: EACL 2026, pages 1940–1954

Comment (Spanish)	Translation (English)	Toxicity Level
Hay que tener fe	You have to have faith	Non-Toxic
Qué mala es la envidia	Envy is such a bad thing	Non-Toxic
Desvergüenza misógina que canta...	Misogynistic shamelessness singing...	Slightly Toxic
Anda que no se les ve el plumero.	It's obvious where they stand	Slightly Toxic
El macho alfa la reclama junto a él en Barcelona.	The alpha male demands her by his side in Barcelona	Toxic
Gobierno Nini.	Useless government	Toxic
Te aviso que tu tienes una única vida y tu amigo más querido también por si no lo sabes	Just reminding you that you have only one life, and so does your dearest friend, in case you forgot	Very Toxic
Otrea inútilbasurilla aspirante a comer de bandeja y gratis (en el talego)	Another useless-trash wannabe freeloading from prison	Very Toxic

Table 1: Examples of Spanish comments with English translations and their toxicity level.

To establish benchmarks, we systematically evaluate three complementary paradigms (Section 4): (i) traditional ML models, (ii) fine-tuned transformer-based encoders, and (iii) prompt-based LLMs. This controlled comparison highlights the trade-offs between efficiency, scale, and domain adaptation in Spanish toxicity detection. Beyond aggregate scores, we conduct detailed error and fairness analyses, showing how entity mentions, dataset skew, and borderline categories systematically shape model predictions.

Our analysis (Section 5) highlights not only performance differences but also linguistic challenges, such as distinguishing sarcasm from non-toxic criticism, and sociopolitical biases, such as the disproportionate toxicity associated with named entities. Through error analysis (Section 6), we identify systematic sources of misclassification and discuss fairness considerations. Our contributions are:

- We release SocialTox, a **new Spanish corpus for toxicity detection** that is both *nuanced*—capturing irony, implicit insults, and borderline categories—and *diverse*, spanning 16 news outlets and multiple sociopolitical domains. To our knowledge, this is the first resource to target such subtle phenomena in Spanish.

- We present a **systematic benchmark** of toxicity detection methods under consistent conditions, comparing ML baselines, encoder-based transformers, and generative LLMs. This enables reproducible evaluation and reveals trade-offs between accuracy, efficiency, and domain adaptation.

- We provide a **comprehensive error and fairness analysis**, showing how entity mentions, annotation ambiguity, and probability miscalibration drive systematic misclassifications. We discuss their implications for the design of robust and trustworthy moderation systems in under-resourced languages.

The dataset and trained models created in this work are publicly available and are intended for research use.

2 Related work

Toxic language is a pervasive issue in online communication. It is broadly defined as language that is offensive, but defining toxicity remains an evolving challenge due to its subjectivity and cultural dependence (Villate-Castillo et al., 2024). Research suggests that toxicity in online discourse extends beyond personal insults to include hate speech, harassment, and misinformation. The importance of addressing online toxicity is underscored by studies highlighting its prevalence and impact (Sheth et al., 2022; Zhao et al., 2021). Automated toxicity detection has thus become a crucial task in NLP to ensure safer online environments while balancing freedom of expression and ethical concerns (Nakov et al., 2021; Zampieri et al., 2020).

A key challenge in building robust toxicity detection systems is the lack of high-quality labelled datasets. Most existing datasets focus on English, such as SFU Opinion and Comments Corpus (SOCC), which contains opinion articles and comments from a Canadian newspaper (Kolhatkar et al., 2020), the semi-Supervised Offensive Language Identification Dataset (SOLID) (Rosenthal et al., 2020), a large-scale dataset with millions of tweets, and the ToxicSpans dataset (Pavlopoulos et al., 2022), which provides whole-post toxicity annotations. In contrast, fewer resources are available for Spanish and other languages. DETOXIS (Detection of Toxicity in Spanish Comments) (Taulé et al., 2021) is one of the few resources for Spanish toxicity detection, focusing on comments related to immigration. NECOS-TOX (News Comments Toxicity Dataset) (López-Úbeda et al., 2024) expands upon this by labelling both toxicity and constructiveness, bridging toxicity detection with discourse quality assessment. Despite the progress in dataset creation, Spanish remains underrepresented in toxicity detection research. To address this gap, our work extends NECOS-TOX by incorporating comments from 16 Spanish news outlets, significantly increasing data diversity.

Some research lines have explored multimodal toxicity datasets. For instance, the DeToxy dataset (Ghosh et al., 2021) comprises spoken utterances annotated for toxicity, while the MuTox dataset (Costa-jussà et al., 2024) offers a multilingual audio-based collection with toxicity labels.

Recent advancements in Artificial Intelligence (AI) and NLP have significantly improved automatic toxicity detection. Traditional ML models such as Support Vector Machines (SVMs) and Naïve Bayes classifiers were gradually replaced by deep learning models, particularly Long Short-Term Memory (LSTM) networks (Zaheri et al., 2020). However, the emergence of Transformer-based models has revolutionized the field, achieving state-of-the-art performance. Models such as BERT and RoBERTa consistently outperform previous approaches in toxicity detection (Villate-Castillo et al., 2024; Pavlopoulos et al., 2021; Taulé et al., 2021) including for Spanish toxic texts (Plaza-del Arco et al., 2021). These findings have been validated in public challenges, such as SemEval 2021 Task 5: Toxic Spans Detection (Pavlopoulos et al., 2021) and Jigsaw Multilingual Toxic Comment Classification (Kivlichan et al., 2020).

While detection is essential, toxicity mitigation and counter-narrative generation are emerging research areas aimed at reducing harm and fostering constructive discussions (Tekiroglu et al., 2022; Rodríguez et al., 2024).

3 Dataset

To overcome the limitations of existing resources, we introduce an extended version of the NECOS-TOX corpus (López-Úbeda et al., 2024). The original dataset contained 1,419 comments from ten news articles published in a single outlet (*El Mundo*)[2]. While valuable as a starting point, its restricted source and size limited both representativeness and linguistic variety. Our extension incorporates 2,592 additional comments sampled from 16 Spanish-language digital news platforms, covering a broader range of topics, stylistic registers, and sociopolitical contexts. The combined resource, named **SocialTox**[3], comprises 4,011 comments annotated for toxicity. To our knowledge, this is the most diverse publicly available Spanish corpus focusing specifically on nuanced toxic language in the news domain. Further details on outlet selection are provided in Appendix A.

3.1 Annotation scheme and validation

Building on prior work (López-Úbeda et al., 2024; Kolhatkar et al., 2020), we refined the annotation scheme[4] to capture both overtly offensive remarks and more subtle forms of toxicity. The categories are defined as follows:

- **Non-Toxic** Comments without toxic content.
- **Slightly Toxic**: Sarcastic, ironic, or mocking comments without constructive criticism, expressing frustration or anger without offensive language. They include rhetorical questions, disparage the content or the author, and may be offensive to some users.
- **Toxic**: Expressions that include mockery, nicknames, or inappropriate jokes intended to generate conflict without resorting to insults. They also include disqualifications, delegitimization, or pejorative comments towards individuals, groups, or institutions, as well as alarmist, denialist, or conspiratorial discourse that denies scientific evidence.
- **Very Toxic**: Comments that include threats, personal attacks, or insults directed at users, authors, public figures, or collectives. They also contain expressions that wish to harm or curse others, as well as denigrating or humiliating comments. In addition, they include those that incite hatred based on nationality,

[2]https://www.elmundo.es/

[3]https://huggingface.co/datasets/gplsi/SocialTOX

[4]https://huggingface.co/datasets/gplsi/SocialTOX/blob/main/Gui%CC%81a_Toxicidad_SocialFairness.pdf

age, gender, sexual orientation, or other aspects of identity, including racist, sexist, homophobic, transphobic, and any form of discrimination.

Three annotators conducted the labelling: two with linguistic expertise and one with a sociology background, ensuring both linguistic sensitivity and awareness of social implications. To assess reliability, we applied Fleiss' kappa (Landis and Koch, 1977). Initial pilot annotations revealed low agreement ($\kappa < 0.30$), primarily due to confusion between the *Toxic* and *Very Toxic* categories. In practice, highly aggressive comments are rare in moderated news forums (only 85 in the original NECOS-TOX), and moderators often remove them before data collection. Following this observation, we consolidated the scheme into three levels (*Non-Toxic*, *Slightly Toxic*, *Toxic*), while retaining examples of *Very Toxic* for documentation. This refinement led to a substantial increase in agreement: κ improved to 0.74 after four iterative calibration rounds. According to the widely used interpretation of Landis and Koch (1977), this constitutes *substantial agreement*. The process illustrates the inherent subjectivity of toxicity annotation, but also the importance of carefully designed guidelines and interdisciplinary adjudication.

3.2 Semi-automatic annotation process

To scale the annotation while preserving quality, we adopted a semi-automatic strategy inspired by the human-in-the-loop (HITL) active learning framework of Botella-Gil et al. (2024). This approach combines model-assisted pre-annotation with systematic human validation, reducing annotation effort without compromising reliability. The process is illustrated in Figure 1. A classifier is trained to distinguish between the three classes in the dataset (Task 1: Train the model). It is then used to automatically label all unannotated comments (Task 2: Predict data), identifying those for which it exhibits the greatest uncertainty. The annotators review these cases (Task 3: Revise) and, after validation, are incorporated into the training partition of the corpus.

The procedure unfolded as follows. First, we created an initial training set that was representative across classes. Since the original NECOS-TOX corpus was limited to a single outlet, we randomly annotated 1,000 additional comments from 16 diverse media platforms. Combined with the 1,419

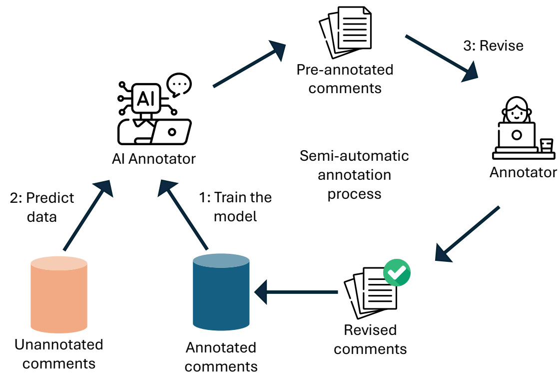


Figure 1: Semi-automatic annotation process with human-in-the-loop active learning.

Sets	Non-Toxic	Slightly Toxic	Toxic	Total
Training	1,052	1,243	748	3,043
Test	381	439	148	968
Total	1,433	1,682	896	4,011

Table 2: Final distribution of comments by class and dataset split.

NECOS-TOX examples, this yielded 2,419 manually labelled instances used to train the first classifier.

The model was then applied to the pool of unlabelled comments. Following standard active learning practice (Monarch, 2021), the system identified cases with the highest prediction uncertainty, which were most informative for model improvement. Annotators reviewed these uncertain samples in batches of 100, correcting labels where necessary. The validated examples were merged into the training set, and the model was retrained. This review-retrain cycle was repeated 16 times, steadily improving both model accuracy and annotation efficiency.

To guarantee unbiased evaluation, we reserved a stratified test set before initiating the active learning phase (Table 2). The final dataset consists of 4,011 comments, with 75% used for training and 25% reserved for testing. As Table 2 shows, the *Slightly Toxic* class is the most frequent, reflecting the prevalence of irony, sarcasm, and indirect disparagement in online news discussions. This category acts as a boundary between *Non-Toxic* and *Toxic* comments, introducing natural ambiguity that poses a significant challenge for automatic classifiers. Rather than being a limitation, this distribution captures a realistic sociolinguistic phenomenon and enables deeper investigation of borderline toxicity.

4 Experimental setup

We evaluate three complementary paradigms for toxicity classification: (i) traditional ML models, (ii) fine-tuned transformer-based encoders, and (iii) prompt-based LLM. To ensure comparability, all models were trained and evaluated under consistent preprocessing, hyperparameter optimization, and evaluation protocols.

Data pre-processing A standardized preprocessing pipeline was applied across all systems to ensure clean and consistent input. The following steps were performed: (a) lowercasing, when required by the target model, (b) removal of URLs, (c) deletion of user mentions (e.g., `@user_1`) to anonymize and reduce noise, (d) removal of thread-specific hashtags containing numbers or bracketed letters, and (e) normalization of spacing.

4.1 Baseline Models: Traditional ML

As an initial baseline, we evaluated traditional ML classifiers using the `PyCaret` framework (Ali, 2020). Logistic regression consistently emerged as the best-performing method across both raw and preprocessed corpora, establishing a reliable lower bound for subsequent transformer-based and LLM-based approaches.

All models were trained under standardized preprocessing and evaluation protocols to ensure fair comparison. Full implementation details, including feature extraction, evaluation settings, and hyperparameter configurations, are reported in Appendix B.

4.2 Fine-Tuning Transformer-Based Encoders

To determine the most suitable models for the toxicity classification task, we fine-tuned a range of transformer-based encoders with different scales, training corpora, and language coverage. This allowed us to establish robust baselines while systematically assessing the impact of model architecture and linguistic scope. We experimented with three representative families: RoBERTuito (Spanish tweets), RoBERTa-bne (curated Spanish corpus), and XLM-RoBERTa (multilingual web-scale data). Details of each model are provided in Appendix C.

Optimization was performed with Bayesian search using `Optuna` (Akiba et al., 2019), targeting macro-averaged F1 (*Macro-F1*) to account for class imbalance. Early stopping was applied with a maximum of 100 epochs. Loss functions were chosen according to task type (binary vs. multi-class). Hyperparameter ranges and final configurations of the best-performing models are reported in Appendix C.

The full ranges of hyperparameters considered, along with the final configurations of the best-performing models, are reported in Appendix C.

4.3 Prompt-Based Classification Using LLM

We further explored prompt-based classification using off-the-shelf LLMs from the Llama family. Specifically, we experimented with two Llama Instruct models of different scales (3B and 8B), chosen for their strong multilingual capabilities and competitive performance on standard benchmarks.

Two evaluation setups were considered: (i) few-shot classification on both binary and multi-class tasks, and (ii) fine-tuning on the same datasets to assess the benefits of task-specific adaptation. In both scenarios, we employed prompts derived from the dataset's annotation guidelines, carefully integrating class descriptions to minimize ambiguity.

This setup enables us to evaluate the out-of-the-box potential of LLMs for toxicity detection and quantify the improvements gained through fine-tuning when compared to traditional ML baselines and transformer-based models. Full model specifications and prompts used are provided in Appendix D.

5 Result and Discussion

This section reports the performance of the proposed models on both binary and multi-class toxicity detection tasks and provides a critical discussion of the findings.

5.1 Binary Classification Results

Table 3 summarizes the results for binary classification. As expected, traditional ML methods establish a weak lower bound (Macro-F1 ≈ 0.62), confirming their limitations in capturing subtle pragmatic and contextual cues. In contrast, fine-tuned transformer-based encoders and LLMs achieve significantly higher performance.

The best-performing model is the fine-tuned Llama-3.1-8B-Instruct (4.3.4), which reaches a Macro-F1 of 0.785. Interestingly, this score is only marginally higher than that of RoBERTa-base-bne (4.2.1, Macro-F1 = 0.768), a compact encoder with just 125M parameters (1.56% of the size of Llama-3.1-8B). This result highlights a central finding: domain-adapted encoder models can rival or

ID	Model	F1 per class Non-Toxic	Toxic	Precision	Recall	F1
4.1. Machine learning						
4.1.1	LogisticRegression	0.529	0.713	0.624	0.620	0.621
4.1.2	LogisticRegression_prep	0.533	0.721	0.631	0.626	0.627
4.2. Fine-Tuning encoders						
4.2.1	RoBERTa-base-bne	<u>0.714</u>	0.821	<u>0.770</u>	<u>0.766</u>	<u>0.768</u>
4.2.2	RoBERTuito-base-uncased	0.712	0.807	0.758	0.761	0.759
4.2.3	RoBERTuitoo-base-cased	0.709	0.804	0.755	0.758	0.756
4.2.4	XLM-RoBERTa-large	0.700	0.817	0.763	0.755	0.759
4.3 Off-the-shelf models						
4.3.1	Llama-3.2-3B-Instruct	0.613	0.336	0.674	0.586	0.475
4.3.2	Llama-3.1-8B-Instruct	0.640	0.519	0.689	0.648	0.579
4.3 Fine-tuning models						
4.3.3	Llama-3.2-3B-Instruct	0.702	<u>0.823</u>	<u>0.770</u>	0.758	0.763
4.3.4	Llama-3.1-8B-Instruct	**0.721**	**0.849**	**0.809**	**0.776**	**0.785**

Table 3: Performance comparison of encoder and generative models in binary classification tasks. The best results for each metric are shown in bold, and the second-best results are underlined.

even surpass billion-parameter LLMs in specialized downstream tasks. Such efficiency has important implications for sustainable and deployable moderation systems, particularly in low-resource settings. Off-the-shelf LLMs evaluated in a 5-shot setting perform poorly (e.g., Llama-3.2-3B-Instruct, F1 = 0.475), often falling below reasonable baselines. Off-the-shelf LLMs were evaluated under 0-, 3-, and 5-shot configurations, with the best results obtained in the 5-shot setting. This confirms that task-specific fine-tuning remains indispensable and suggests that Spanish toxicity detection is still underrepresented by current multilingual pre-training.

5.2 Multi-Class Classification Results

Table 4 reports results for the more challenging three-class setup. Performance drops considerably across all models: the best model, Llama-3.1-8B-Instruct (4.3.8), reaches only 0.633 Macro-F1, while RoBERTa-base-bne (4.2.5) achieves 0.610. This represents a performance gap of more than 0.15 compared to binary classification.

The main bottleneck lies in the *Slightly Toxic* category, which exhibits substantial confusion with both Non-Toxic and Toxic classes. This finding confirms our annotation analysis: *borderline or implicit toxicity is inherently ambiguous and difficult to operationalize*. Notably, even with careful guidelines, the model struggles to separate irony and sarcasm from legitimate criticism. Future annotation efforts should therefore consider richer context (e.g., conversation threads) or multi-label setups to better capture subtlety.

Traditional models again perform poorly (Macro-F1 $\approx$ 0.43), particularly in the Toxic class (F1 $<$ 0.27), underscoring their inability to capture nuanced linguistic phenomena.

5.3 Key Observations

Our experiments yield three main takeaways:

- **Efficiency vs. scale:** Compact encoders for Spanish language (e.g., RoBERTa-base-bne) rival instruction-tuned LLMs when adapted to the domain, showing that *domain adaptation often outweighs sheer scale*.
- **Ambiguity of borderline cases:** Performance drops sharply in the multi-class setting, confirming that *slightly toxic* is difficult to separate and that toxicity behaves more like a continuum than a set of discrete categories.
- **Few-shot limitations:** Multilingual LLMs underperform without adaptation, underscoring that plug-and-play moderation is unrealistic for Spanish and similar under-resourced languages.

Together, these results stress the dual importance of **nuanced annotation** and **efficient domain-specific modeling**. For transparency and reproducibility, we release the best-performing models:

- 4.2.1 RoBERTa-base-bne[5] (Binary)
- 4.3.2 Llama-3.1-8B-Instruct[6] (Binary)

[5] https://huggingface.co/gplsi/Toxicity_model_binary

[6] https://huggingface.co/gplsi/Toxicity_model_Llama_3.1_8B_binary

ID	Model	F1 per class Non-Toxic	Slightly Toxic	Toxic	Precision	Recall	F1
4.1. Machine learning							
4.1.3	LogisticRegression	0.503	0.525	0.246	0.425	0.427	0.425
4.1.4	LogisticRegression_prep	0.518	0.531	0.265	0.438	0.442	0.438
4.2. Fine-Tuning encoders							
4.2.5	RoBERTa-base-bne	<u>0.724</u>	<u>0.665</u>	<u>0.450</u>	<u>0.617</u>	<u>0.609</u>	<u>0.613</u>
4.2.6	RoBERTuito-base-uncased	0.696	0.643	0.386	0.588	0.568	0.575
4.2.7	RoBERTuito-base-cased	0.669	0.628	0.430	0.582	0.571	0.576
4.2.8	XLM-RoBERTa-large	0.703	0.529	0.396	0.545	0.560	0.543
4.3 Off-the-shelf models							
4.3.5	Llama-3.2-3B-Instruct	0.402	0.463	0.140	0.297	0.229	0.251
4.3.6	Llama-3.1-8B-Instruct	0.656	0.602	0.251	0.514	0.499	0.503
4.3 Fine-tuning models							
4.3.7	Llama-3.2-3B-Instruct	0.677	0.593	0.419	0.567	0.562	0.563
4.3.8	Llama-3.1-8B-Instruct	**0.732**	**0.686**	**0.480**	**0.645**	**0.627**	**0.633**

Table 4: Performance comparison of encoder and generative models in multi-class classification tasks. The best results for each metric are shown in bold, and the second-best results are underlined.

- 4.2.5 RoBERTa-base-bne[7] (Multi-class)
- 4.3.6 Llama-3.1-8B-Instruct[8] (Multi-class)

6 Error Analysis

To move beyond aggregate metrics, we conducted an error analysis to uncover systematic factors driving misclassifications. We focus on: (i) class-level confusions and calibration, examining linguistic phenomena behind errors; and (ii) entity-related biases affecting fairness.

6.1 Entity Bias and Anonymization

We observed that a recurring pattern is the strong association of political entities with *Toxic* labels. Mentions of specific figures or organizations can act as shortcuts, where models flag entities as inherently toxic rather than relying on genuine linguistic cues. To quantify this, we anonymized high-confidence (≥ 0.97) PERSON and ORG mentions using placeholders ([PER], [ORG]) via BSC-LT's multilingual anonymization model[9]. Table 5 reports the results. Across both binary and multi-class settings, anonymization consistently reduced predictive performance. For instance, RoBERTa-base-bne drops from 0.768 to 0.758 Macro-F1 in binary classification, while the Toxic class in the multi-class setup falls from 0.450 to 0.378. These findings confirm that entity mentions serve as weak but effective proxies for toxicity, artificially boosting performance in non-anonymized settings.

Statistical testing further corroborates the presence of entity bias. A Chi-squared test over the training data yields $\chi^2 = 2906.272$ with dof $= 2726$ and $p = 0.00823$, while Cramér's V = 0.771, indicating a strong, non-random association between entity mentions and toxicity labels. A detailed breakdown at the level of individual leaders and parties is provided in Appendix E.1. These results underscore the necessity of integrating debiasing techniques (e.g., anonymization, counterfactual data augmentation, or entity-balancing) into future toxicity detection pipelines.

6.2 Confusion Across Toxicity Levels

To better understand misclassification patterns, we conducted pairwise binary classification experiments with RoBERTa-base-bne, selected for its efficiency and competitive performance. Each experiment isolated two categories, enabling a clearer view of class boundaries. The settings were: (a) *Slightly Toxic* vs. *Toxic*; (b) *Non-Toxic* vs. *Toxic*, excluding *Slightly Toxic*; and (c) *Non-Toxic* vs. *Slightly Toxic*, excluding *Toxic*. Table 6 reports the results.

To complement the quantitative results, we manually inspected misclassified instances from the *Non-Toxic vs. Toxic* setup (Table 7). False positives reveal the model's oversensitivity to emotionally charged expressions that are critical or sarcastic but not abusive (e.g., "unbearable" in ID 2). False negatives, in contrast, often involve implicit aggression or veiled threats that require pragmatic or cultural knowledge (e.g., "they are rats already" in ID 3).

[7] https://huggingface.co/gplsi/Toxicity_model/

[8] https://huggingface.co/gplsi/Toxicity_model_Llama_3.1_8B

[9] https://huggingface.co/BSC-LT/roberta_model_for_anonimization

Model	F1 per class			Precision	Recall	F1
	Non-Toxic	Slightly Toxic	Toxic			
Binary classification						
RoBERTa-base-bne	0.716	-	0.799	0.755	0.763	0.758
RoBERTuito-base-cased	0.706	-	0.801	0.752	0.756	0.753
XLM-RoBERTa-large	0.702	-	0.800	0.749	0.753	0.751
Multi-class classification						
RoBERTa-base-bne	0.711	0.613	0.378	0.565	0.570	0.567
RoBERTuito-base-cased	0.676	0.583	0.422	0.555	0.569	0.560
XLM-RoBERTa-large	0.695	0.551	0.391	0.541	0.555	0.545

Table 5: Performance comparison of encoder models (Section 4.2) on binary and multi-class classification tasks, applying entity anonymization to the dataset.

Experiment	F1 per class			Precision	Recall	F1
	Non-Toxic	Slightly Toxic	Toxic			
Slightly Toxic vs Toxic	–	0.803	0.439	0.619	0.624	0.621
Non-Toxic vs Toxic	0.879	–	0.704	0.786	0.798	0.791
Non-Toxic vs Slightly Toxic	0.696	0.754	–	0.728	0.724	0.725

Table 6: Performance metrics for pairwise binary classification experiments using RoBERTa-base-bne.

ID	Label	Prediction	Text	Translation
1	0	1	*hay que dejar de utilizar las palabras ,fascista y comunista ,un sintendido en la sociedad actual*	People should stop using the words ‘fascist’ and ‘communist’, as they are nonsensical in today’s society.
2	0	1	*de acuerdo con las y los expertos. insufrible*	I agree with the experts. Unbearable.
3	1	0	*sería lo más práctico, porque para empezar ratas ya son*	That would be the most practical thing, since they are rats already anyway.
4	1	0	*te aviso que tu tienes una unica vida y tu amigo mas querido tambien por si no lo sabes*	Just a reminder: you only have one life, and so does your dearest friend, in case you didn’t know.

Table 7: Examples of misclassified instances in the Non-Toxic vs Toxic binary classification experiment using RoBERTa-base-bne.

We computed Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) for each class. Calibration analysis (Appendix E.2) shows *Non-Toxic* predictions are well-calibrated (ECE = 0.091), *Slightly Toxic* moderately (ECE = 0.122), and *Toxic* substantially underconfident (ECE = 0.218, MCE = 0.461). These findings reinforce our previous observation: while class separability already presents challenges, probability estimates themselves are unreliable, particularly for the most socially critical category.

7 Conclusions

This work introduces a new resource and benchmark for Spanish toxicity detection, filling a longstanding gap in multilingual moderation research. By expanding NECOS-TOX with comments from 16 diverse news outlets, refining the annotation protocol, and adopting a semi-automatic active learning pipeline, we provide a corpus that captures not only overt insults but also subtle forms of toxicity such as sarcasm, delegitimization, and veiled threats.

Our experiments show that, for Spanish language, carefully fine-tuned transformer encoders can rival billion-parameter LLMs, highlighting the competitiveness of efficient, domain-adapted models for socially relevant NLP tasks. At the same time, performance drops in multi-class settings confirm the intrinsic difficulty of distinguishing borderline cases like *Slightly Toxic* from adjacent categories. This challenge is both technical and conceptual, pointing to the need for richer annotation schemes and context-sensitive modeling approaches.

From an ethical perspective, our analysis reveals strong associations between named entities and toxicity predictions, raising fairness concerns. While anonymization reduces such biases at a modest cost in accuracy, it aligns with European regulatory frameworks (e.g., the EU AI Act) that prioritize trustworthy AI.

Future Work A first step is to expand the dataset in scale and topical coverage, for instance by including additional news outlets. Another priority is to refine the annotation guidelines to better capture borderline cases and to explore whether ordinal or multi-label schemes reduce ambiguity in categories such as *Slightly Toxic*. Building on our error analysis, incorporating limited conversational context could help models distinguish irony from legitimate criticism. Finally, fairness concerns observed in entity bias motivate the development of lightweight anonymization or debiasing strategies that preserve linguistic nuance while mitigating shortcuts.

Limitations

Although this work makes important contributions, several limitations should be acknowledged. First, the dataset remains modest in size compared to large-scale resources available for English. While our focus on diversity and nuanced cases (e.g., irony, sarcasm) enriches the corpus, the limited scale constrains generalization and may leave rare but socially critical phenomena underrepresented.

Second, despite the iterative annotation protocol and substantial agreement ($\kappa = 0.74$), toxicity judgments are inherently subjective. The consolidation of the *Very Toxic* category into *Toxic* facilitated higher reliability but reduced granularity, potentially obscuring distinctions between severe abuse (e.g., direct threats) and milder derogatory language.

Third, all comments were annotated in isolation. Many subtle cases such as sarcasm, veiled threats, or delegitimization, depend on conversational or pragmatic context that is absent in our dataset. Incorporating thread-level or user-level context may help resolve such ambiguities.

Finally, our error analysis revealed entity-related biases: models often rely on named entities as shortcuts to predict toxicity. While anonymization mitigates this effect, we did not implement systematic debiasing techniques.

Overall, these limitations do not undermine the main contributions of this work but rather point to directions for extending and strengthening research on nuanced toxicity detection in under-resourced languages.

Ethical Considerations

This work uses and releases a dataset of user-generated comments collected from publicly accessible news platforms. Given the sensitive nature of toxic language, the data may contain offensive expressions. All comments were processed prior to use and release, and explicit personally identifying information was removed where present. The dataset and trained models are released strictly for research purposes, in accordance with the original access conditions of the data.

Toxicity annotation is inherently subjective and context-dependent. To mitigate individual bias, annotations were performed by multiple annotators following detailed guidelines, with iterative calibration to improve consistency. Nevertheless, some ambiguity remains, particularly for borderline cases such as sarcasm or implicit aggression.

Our analysis also reveals that models may rely on named entities as shortcuts for toxicity prediction, raising fairness concerns. We therefore examine entity-related biases and assess the impact of anonymization as a mitigation strategy. The released resources are intended to support research on trustworthy and responsible toxicity detection and should not be deployed in real-world moderation settings without further ethical and legal assessment.

Acknowledgments

This work is funded by the *Ministerio para la Transformación Digital y de la Función Pública* and *Plan de Recuperación, Transformación y Resiliencia*, funded by the European Union, NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*. This work has also been partially supported by Project COOLANG.CONSENSO/TRIVIAL (PID2021-122263OB-C21/PID2021-122263OB-C22), HEART-NLP.UJA/UA (PID2024-156263OB-C21/PID2024-156263OB-C22), VERITAS-H (AIA2025-163322-C64), ROMANET (CERV-2024-CHAR-LITI-101215052)

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Moez Ali. 2020. PyCaret: An open source, low-code machine learning library in Python. PyCaret version 1.0.0.

Beatriz Botella-Gil, Robiert Sepúlveda-Torres, Alba Bonet-Jover, Patricio Martínez-Barco, and Estela Saquete. 2024. Semi-automatic dataset annotation applied to automatic violent message detection. IEEE Access.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. CoRR, abs/1911.02116.

Marta R Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. 2024. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. arXiv preprint arXiv:2401.05060.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. Procesamiento del Lenguaje Natural, 68.

Sreyan Ghosh, Samden Lepcha, Sahni Sakshi, Rajiv Ratn Shah, and Srinivasan Umesh. 2021. Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances. arXiv preprint arXiv:2110.07592.

Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw multilingual toxic comment classification. Kaggle.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. Corpus pragmatics, 4:155–190.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. biometrics, pages 159–174.

Pilar López-Úbeda, Flor Miriam Plaza-del Arco, Manuel-Carlos Díaz-Galiano, and M Teresa Martín-Valdivia. 2024. Toxicity in spanish news comments and its relationship with constructiveness. Procesamiento del lenguaje natural, 73:43–53.

R Monarch. 2021. Munro. active learning and annotation for human-centered ai.

Preslav Nakov, Vibha Nayak, Kyle Dent, Ameya Bhatawdekar, Sheikh Muhammad Sarwar, Momchil Hardalov, Yoan Dinkov, Dimitrina Zlatkova, Guillaume Bouchard, and Isabelle Augenstein. 2021. Detecting abusive language on online platforms: A critical analysis. arXiv preprint arXiv:2103.00153.

John Pavlopoulos, Léo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)., Dublin, Ireland. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pages 59–69.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 7235–7243, Marseille, France. European Language Resources Association.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Urena-López, and M Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. Expert Systems with Applications, 166:114120.

María Estrella Vallecillo Rodríguez, Maria Victoria Cantero Romero, Isabel Cabrera De Castro, Arturo Montejo Ráez, and María Teresa Martín Valdivia. 2024. Conan-mt-sp: A spanish corpus for counternarrative using gpt models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3677–3688.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. Solid: A large-scale semi-supervised dataset for offensive language identification. arXiv preprint arXiv:2004.14454.

Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. Neurocomputing, 490:312–318.

Mariona Taulé, Alejandro Ariza, Montserrat Nofre, Enrique Amigó, and Paolo Rosso. 2021. Overview of detoxis at iberlef 2021: detection of toxicity in comments in spanish. Procesamiento del lenguaje natural, 67:209–221.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. arXiv preprint arXiv:2204.01440.

Guillermo Villate-Castillo, Javier Del Ser, and Borja Sanz Urquijo. 2024. A systematic review of toxicity in large language models: Definitions, datasets, detectors, detoxification methods and challenges.

Sara Zaheri, Jeff Leath, and David Stroud. 2020. Toxic comment classification. SMU Data Science Review, 3(1):13.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). arXiv preprint arXiv:2006.07235.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. A comparative study of using pre-trained language models for toxic comment classification. In Companion Proceedings of the Web Conference 2021, pages 500–507.

A Source of user comments

The sources used to collect user comments were selected exclusively from Spanish online newspapers. The selection aimed to include outlets that actively host user discussions and maintain publicly accessible comment sections. We also sought to balance sources in terms of editorial style and audience, combining newspapers recognized for their journalistic quality with others characterized by more sensationalist or opinion-driven content. Table 8 lists the sixteen outlets included in the dataset.

Newspaper	URL
Okdiario	`https://okdiario.com/`
Marca	`https://www.marca.com/`
El Mundo	`https://www.elmundo.es/`
El País	`https://elpais.com/`
NTV España	`https://ntvespana.com/`
The Objective	`https://theobjective.com/`
Alerta Digital	`https://www.alertadigital.com/`
El Debate	`https://www.eldebate.com/`
Vozpópuli	`https://www.vozpopuli.com/`
El Español	`https://www.elespanol.com/`
ABC	`https://www.abc.es/`
El Confidencial	`https://www.elconfidencial.com/`
El Diario	`https://www.eldiario.es/`
Moncloa	`https://www.moncloa.com/`
Periodista Digital	`https://www.periodistadigital.com/`
20 Minutos	`https://www.20minutos.es/`

Table 8: Sources of user comments: Spanish online newspapers. Last accessed: 8 January 2026.

B Traditional Machine Learning Baselines

This appendix provides a detailed account of the setup and configurations used for traditional ML baselines.

B.1 Feature Extraction

Texts were vectorized using a TF-IDF representation restricted to the 5,000 most frequent terms. This dimensionality limit balanced computational efficiency with representational coverage. The resulting sparse matrix was used as input for all classifiers tested.

B.2 Experimental Settings

To ensure comparability across models, the following settings were applied to every experiment:

- **Normalization:** Min-Max scaling applied to feature values.
- **Cross-validation:** Stratified k-fold splits to preserve class distributions.
- **Imbalance handling:** Automatic reweighting or resampling strategies provided by `PyCaret`.
- **Evaluation metric:** Macro-F1, chosen for its sensitivity to imbalanced and multi-class scenarios.
- **Reproducibility:** Random seed fixed to 42.

B.3 Optimal Model and Hyperparameters

Across all tested algorithms, logistic regression consistently yielded the highest performance. The final configuration was:

- Regularization strength $C = 1.0$
- L2 penalty
- `lbfgs` solver
- Automatic multi-class handling
- Maximum of 1000 iterations

C Transformer Encoder Details

This appendix provides the full details of the hyperparameter exploration and the final configurations of the top-performing models.

C.1 Pre-trained Models

- **RoBERTuito-base** (cased/uncased): trained on 500M Spanish tweets, designed for user-generated content (Pérez et al., 2022).
- **RoBERTa-base-bne**: part of the MarIA Project, trained on the curated Spanish corpus of the National Library of Spain (Fandiño et al., 2022).
- **XLM-RoBERTa-large**: a multilingual encoder trained on CommonCrawl, covering 100+ languages (Conneau et al., 2019).

C.2 Hyperparameter Ranges

Table 9 presents the ranges of hyperparameters explored during the fine-tuning process. These values were chosen to cover a wide spectrum of possible configurations, balancing commonly used defaults with more extreme values.

Hyperparameter	Range Tested
Learning rate	1×10^{-6} to 1×10^{-4}
Batch size	4, 8, 16, 32
Weight decay	1×10^{-12} to 1×10^{-1}
Adam epsilon	1×10^{-10} to 1×10^{-6}
Dropout (multi-class only)	{0.1, 0.2, 0.25, 0.3}

Table 9: Hyperparameter ranges explored during fine-tuning.

Overall, RoBERTuito-base and RoBERTa-base-bne achieved competitive results in the binary setting, while XLM-RoBERTa-large consistently benefited the most from larger batch sizes and lower learning rates.

C.3 Final Configurations of Top Models

Table 10 summarizes the specific hyperparameter values corresponding to the top-performing models across both binary and multi-class settings. Each configuration represents the best trial found within the 20 optimization runs carried out for each model. These results highlight that optimal configurations varied not only across model families, but also between binary and multi-class setups.

D LLM Model Details and Prompts

D.1 Llama Model Specifications

We experimented with the following open-weight instruction-tuned models:

- **Llama-3.2-3B Instruct**[10]: Multilingual dialogue-oriented model with 128k context window, fine-tuned and aligned with human feedback.
- **Llama-3.1-8B Instruct**[11]: Optimized for general-purpose dialogue, with an 8k context window and human-aligned fine-tuning.

D.2 Prompt Design

Prompts were adapted from the dataset annotation guidelines. The system instruction outlined the classification task and provided concise definitions of the three classes (*non-toxic*, *slightly toxic*, *toxic*). Figure 2 shows the English version of the multi-class prompt template, originally written in Spanish, used for instruction tuning. In the few-shot setting, five sample instances per class were added to the prompt.

E Detailed Error Analysis

E.1 Detailed Analysis of Entity Bias

To dig deeper into entity bias, we analyzed the toxicity distribution among the main Spanish political parties (*PP*, *VOX*, *PODEMOS*, *PSOE*) and a selection of their representatives (*Feijóo*, *Abascal*, *Iglesias*, *Sánchez*), shown in Figure 3. The charts reveal that the proportions of *Toxic* and *Slightly Toxic* labels are high and broadly similar across both parties and their leaders, with some variations: parties *PSOE* and *VOX* have a higher share of mentions labeled as toxic in comparison to *PP* and *PODEMOS*, while among representatives, *Sánchez* and *Abascal* exhibit relatively greater toxicity than *Feijóo* or *Iglesias*.

To quantify the association between entities and toxicity labels, we constructed a contingency table for four major Spanish parties and their leaders (Table 11). A Chi-squared test confirms statistical significance ($\chi^2 = 26.0, p < 0.01$), though the effect size is modest (Cramér's V = 0.065), indicating a dataset-level label skew rather than a dominant effect.

This asymmetry raises two concerns. First, annotators' political orientations may unconsciously influence label assignments, particularly in borderline or sarcastic comments. Second, it reveals a dataset-level *label bias*: toxicity distributions reflect the political salience of entities rather than intrinsic linguistic properties. Without explicit debiasing, models trained on such skewed data risk over-penalizing discourse about certain actors, encoding political prejudice into automated moderation pipelines. These findings motivate the use of entity anonymization and other counterfactual or debiasing strategies to mitigate shortcut learning, as discussed in the main text (Section 6.1).

E.2 Confidence and Calibration Analysis

Figure 4 complements the quantitative metrics in Section 6.2. The top row shows the predicted probability distributions per class, while the bottom row presents the corresponding calibration curves.

[10]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[11]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

ID	Classification	Model	Epochs	L. Rate	Train BS	WD	Adam	Dropout
4.2.1	Binary	RoBERTa-base-bne	10	2.45E-06	16	0.00E-00	1.00E-08	-
4.2.2	Binary	RoBERTuito-base-uncased	7	1.10E-05	8	6.37E-03	3.15E-08	-
4.2.3	Binary	RoBERTuito-base-cased	19	1.91E-06	32	1.80E-05	1.94E-07	-
4.2.4	Binary	XLM-RoBERTa-large	9	3.30E-06	32	1.08E-12	4.30E-09	-
4.2.5	Multi-class	RoBERTa-base-bne	7	1.51E-06	16	3.88E-12	2.80E-08	-
4.2.6	Multi-class	RoBERTuito-base-uncased	11	2.77E-06	8	1.11E-06	6.12E-07	0.25
4.2.7	Multi-class	RoBERTuito-base-cased	10	3.44E-06	16	3.63E-10	2.77E-10	-
4.2.8	Multi-class	XLM-RoBERTa-large	7	1.86E-06	4	6.87E-04	2.75E-10	0.20

Table 10: Final hyperparameter configurations of top-performing models.

Entity	0 (Non-Toxic)	1 (Slightly)	2 (Toxic)
None	968	1,127	657
PODEMOS	20	27	11
PP	21	24	15
PSOE	24	53	48
VOX	19	12	17

Table 11: Contingency table of toxicity labels for selected entities (training set).

Class	ECE	MCE	Avg Pred	Actual Frac
Non-Toxic	0.091	0.192	0.410	0.394
Slightly Toxic	0.122	0.224	0.423	0.454
Toxic	0.218	0.461	0.166	0.153

Table 12: Calibration metrics for RoBERTa-base-bne in multi-class classification. ECE: expected calibration error, MCE: maximum calibration error, Avg Pred: average predicted probability, Actual Frac: empirical fraction of examples.

As shown in Figure 4a, *Non-Toxic* exhibits a bimodal distribution, indicating good separability. *Slightly Toxic* overlaps with neighboring classes, explaining frequent misclassifications. The *Toxic* class is concentrated at low-confidence predictions, consistent with its higher calibration error.

Calibration curves (Figure 4b) corroborate these trends. *Non-Toxic* predictions are close to the diagonal but slightly underconfident at mid-range probabilities. *Slightly Toxic* predictions are overconfident between 0.5-0.8, whereas *Toxic* predictions systematically underestimate confidence. Overall, these results indicate that class imbalance and pragmatic ambiguity contribute to miscalibration, complicating accurate toxicity detection.

System:

You are an expert in detecting toxicity in comments, and your goal is to classify comments based on their level of toxicity. The comments were made on news articles. The toxicity categories are:

Slightly toxic: Comments that express sarcasm, irony, or rhetorical questions, but do not directly attack or degrade others.

Toxic: Comments that contain derogatory or pejorative language, inappropriate jokes, fearmongering, denial of facts, threats, personal attacks, insults, degradation, or racist or sexist language. Only classify a comment as "toxic" if it contains clear attack language, direct insults, or demeaning references.

Non-toxic: Neutral or critical comments that do not include *Toxic* or *Slightly toxic* elements. Note that negative or critical comments (those with a serious or discontented tone) are *Non-toxic* or *Slightly toxic* unless they meet the criteria of the categories above.

Please write the corresponding category immediately after the word "answer." In case of doubt between two labels, choose the one with the lowest or no toxicity level.

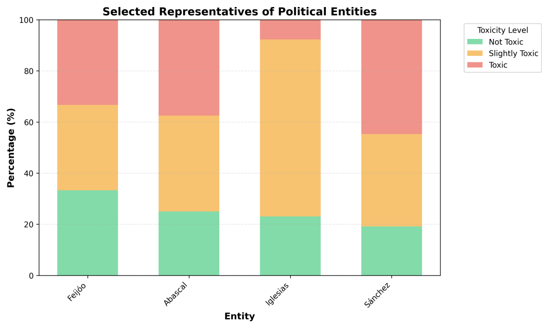
User:

Text: "Narco-Bolivarian Communism"
Please classify this comment, answer:

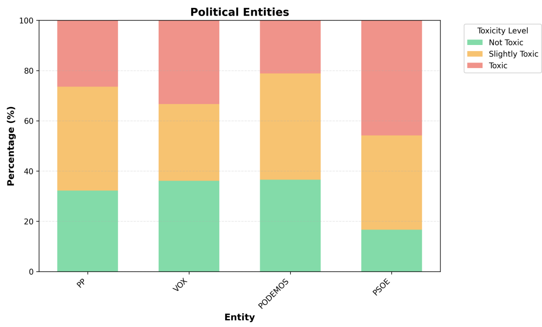
Assistant:

Toxic

Figure 2: English version of the multi-class prompt (originally written in Spanish) used for instruction tuning.

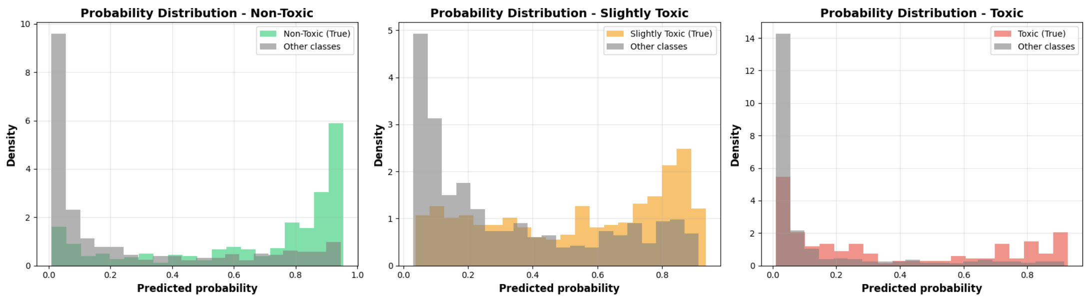


(a) Selected Representatives of Political Entities

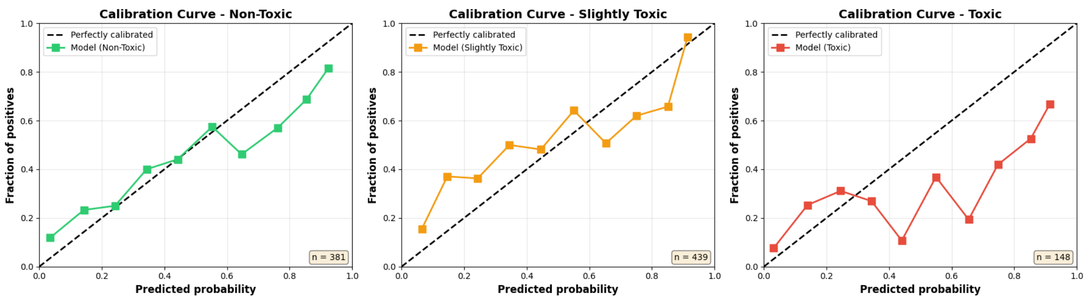


(b) Political Entities

Figure 3: Relative Toxicity Distribution for Political Entities.



(a) Predicted probability distributions per class.



(b) Class-wise calibration curves.

Figure 4: Visualization of predicted probabilities and calibration for RoBERTa-base-bne in multi-class toxicity classification.