informatics

MDPI

Article

Inappropriate Content Classification Model for Digital Violence Detection Using Hybrid Data

Patricio Xavier Zambrano Rodríguez *, Marco Polo Sánchez Aguayo, Carlos Eduardo Anchundia Valencia, Johan Sebastian Illicachi Manzano, Andrea Damarys Oña Calahorrano, Adrian Esteban Paguay Montenegro and Juan Sebastián León Espinosa

Escuela Politécnica Nacional, Facultad de Ingeniería de Sistemas, Quito 170525, Ecuador; marco.sanchez01@epn.edu.ec (M.P.S.A.); carlos.anchundia@epn.edu.ec (C.E.A.V.); johan.illicachi@epn.edu.ec (J.S.I.M.); andrea.onacalahorrano@epn.edu.ec (A.D.O.C.); adrian.paguay@epn.edu.ec (A.E.P.M.); juan.leon01@epn.edu.ec (J.S.L.E.)

* Correspondence: patricio.zambrano@epn.edu.ec; Tel.: +593-2-297-6300

Abstract

This study presents an inappropriate-content classification framework for digital-violence detection in Ecuadorian Spanish, addressing extreme class imbalance and dialectal variation. Data were collected in a post-API setting through a Selenium-based scraping pipeline that reconstructs conversational context using a window of ? = 3 prior interventions. An initial zero-shot labeling attempt with LLMs (Hermes) revealed severe cultural misinterpretation, overestimating the Violence class by 50 times (97.5% false alerts), which motivated full human validation and targeted data engineering. To correct imbalance without contaminating evaluation, the corpus was split before augmentation (70/15/15), and minority classes were selectively leveled via few-shot generation with LLaMA 3.1, followed by strict deduplication and cosine-similarity filtering (? = 0.85) to preserve semantic diversity. Model selection compared BETO and mBERT, with BETO outperforming. Across four training scenarios, naïve oversampling produced artificially inflated metrics indicative of overfitting, whereas the proposed cost-sensitive and regularized configuration (BETO with semantic deduplication and weighted loss) achieved 94.39% accuracy, 0.9429 weighted F1, and 0.9022 macro F1, significantly improving recovery of critical classes. Results highlight that hybrid data are effective only when carefully curated and paired with leakage-free evaluation protocols. This work demonstrates how machine learning innovation and knowledge extraction from heterogeneous data can be combined to build robust models for digital-violence detection in low-resource, culturally specific contexts.

Keywords: digital violence; offensive language detection; natural language processing; Transformer models; fine-tuning; synthetic data; data augmentation; class imbalance; Latin American Spanish; BETO

Check for updates

Received: 23 July 2026
Revised: 4 September 2026
Accepted: 8 September 2026
Published: 16 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Global hyperconnectivity has expanded the exposure surface to cyber threats [1], among which digital violence stands out due to its direct impact on individuals' psychological well-being and reputational integrity, as well as on the governance of communities in social platforms. In this work, we adopt an inappropriate-content classification perspective oriented toward identifying hostile content in Spanish, considering gender-based violence as a relevant—but not exclusive—subset within a broader spectrum of harassment behaviors that includes workplace mobbing dynamics [2,3].

A preliminary analysis of the phenomenon in social media suggests that the main challenge is not only the presence of explicit insults, but the combination of conversational context, irony, and the use of regional idioms that characterizes Latin American settings. In the Ecuadorian context in particular, aggression may manifest through implicit linguistic phenomena (such as irony, sarcasm, or local idioms) that are not captured by keyword-based rules [4]. Consequently, generic filtering systems tend to underestimate these attacks, especially when they rely on superficial lexical cues. This limitation is further exacerbated in Spanish by the lack of public datasets that capture dialectal variants and vernacular sociocultural scenarios—a gap widely recognized in the literature and validated by comparative evaluations across different variants of the language [5–7].

However, attempting to remedy this lack through direct data mining introduces a new technical challenge: the extreme class imbalance inherent to social media streams [5,8]. Because most everyday traffic is non-offensive, instances of digital violence constitute a minority fraction. This asymmetry induces the accuracy paradox [9], in which a classifier may report high overall metrics while failing to recover critical classes (false negatives), an unacceptable risk in content moderation and cybersecurity applications.

Addressing this problem requires a strategy that increases the representativeness of minority categories without degrading the validity of the evaluation set. In this study, we motivate the use of LLM-based synthetic data as a mechanism for controlled class leveling [10]. Recent research suggests that, unlike traditional lexical substitution methods, generative augmentation with large-scale models can mitigate the risk of overfitting by multiplying the volume of information. This significantly improves semantic richness—a property we explicitly ensure through cosine-similarity filtering—and can enhance generalization in hate-recognition tasks [11].

Contributions. The main contributions of this paper are as follows:

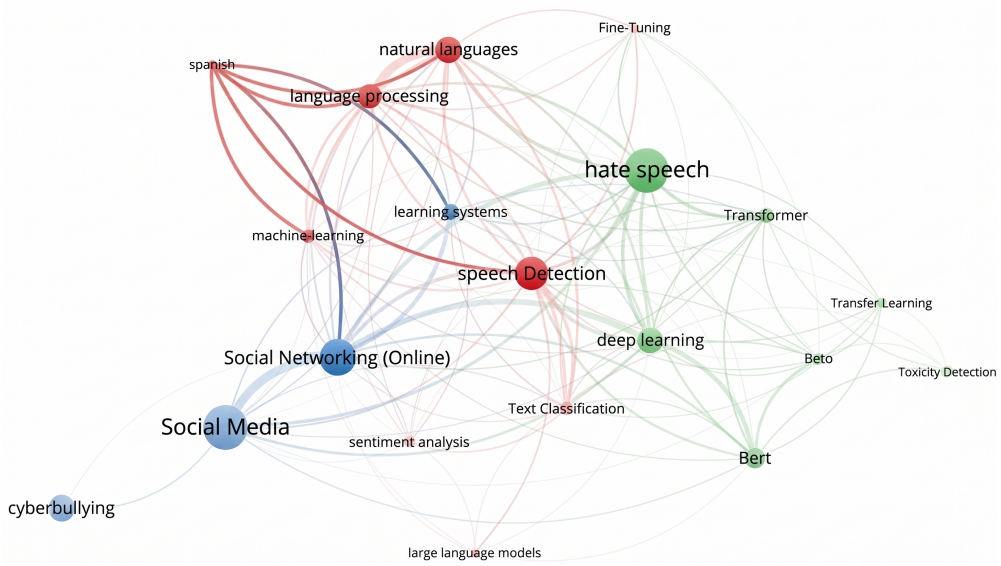
- We construct and curate a corpus of Ecuadorian Spanish social-media conversations for inappropriate-content detection, including conversational context via a fixed window ($k = 3$).
- We propose a controlled synthetic data augmentation protocol (seed selection, few-shot generation, exact deduplication, and cosine-similarity filtering) to mitigate extreme class imbalance while preventing evaluation leakage.
- We evaluate monolingual (BETO) and multilingual (mBERT) Transformer baselines and demonstrate the impact of imbalance-handling strategies (original, undersampling, oversampling, and cost-sensitive training) on per-class recovery.
- We report a cost-sensitive BETO configuration that improves minority-class detection under a strict train/validation/test split and provides detailed error analysis (confusion matrices, ROC, and Precision–Recall curves).

As background, our prior work [12] established an initial methodological workflow for behavioral analysis. Building on that foundation, the present study focuses on developing a robust classification framework tailored to Ecuador's linguistic particularities. To this end, we use the corpus curated through our augmentation strategy to conduct a comparative fine-tuning analysis between two state-of-the-art models: BETO [13] (monolingual) and mBERT [14] (multilingual). The goal is to determine which model better leverages the semantic richness of synthetic data to maximize the recovery of digital violence in dialect-specific environments.

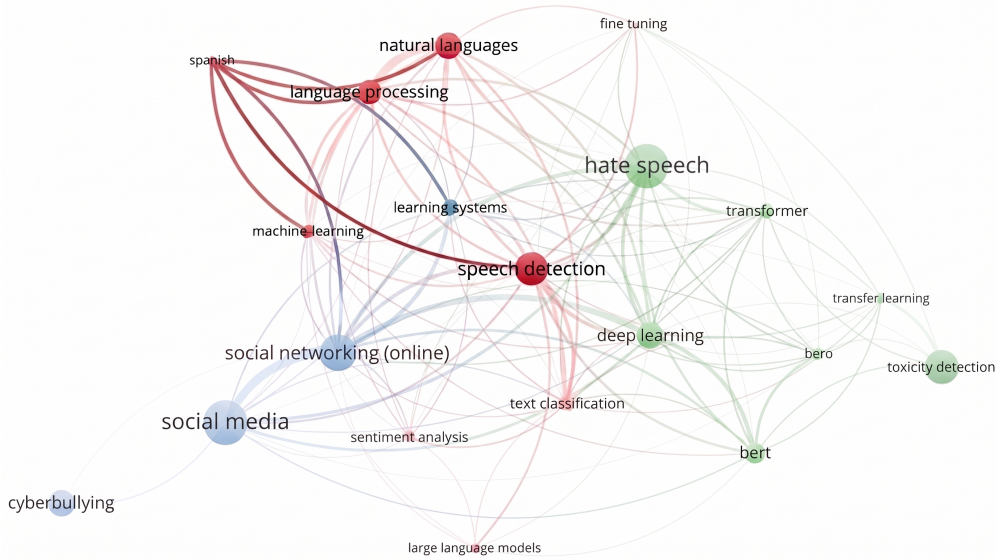
2. Background (State of the Art)

Automatic detection of digital violence in Spanish, particularly in Latin American variants, faces a structural gap, a situation that is visually evidenced by the bibliometric analysis of recent literature (see Figure 1). As observed in the co-occurrence network,

anglocentric terms (e.g., hate speech, deep learning) occupy a central and dominant position, whereas the Spanish node exhibits a lower connection density compared to anglocentric terms, as visualized in the co-occurrence network (Figure 1b). This disparity translates into a scarcity of representative annotated corpora and limited transferability of models trained in other domains/variants. Although there is literature and resources for cyberbullying and offensive language, the evidence shows that performance can vary substantially across Spanish variants and that systems tend to degrade when confronted with local slang and implicit pragmatic phenomena [5,7].



(**a**) Overall view of the network



(**b**) Detail of the Spanish node

Figure 1. Bibliometric analysis evidencing the gap for the Spanish language.

This shortfall is exacerbated by a paradigm shift in data access: the "APIcalypse" has restricted large-scale academic collection, forcing post-API methodologies based on extraction and reconstruction of conversational context [15]. Even when data are available, the natural flow of social media imposes an extreme imbalance that makes aggregated metrics misleading and increases the risk of false negatives in critical classes [9]. This work addresses these limitations through an acquisition pipeline for Ecuadorian Spanish,

controlled synthetic augmentation, and a monolingual discriminative architecture adapted via fine-tuning. Main limitations in related work and our approach. Prior studies often (i) rely on datasets dominated by non-offensive content, reporting aggregate accuracy that hides false negatives in minority classes; (ii) ignore conversational context, which is crucial for implicit aggression (irony, sarcasm, dark humor); and (iii) lack dialect-specific resources for Latin American Spanish, limiting cross-variant transfer. We address these limitations by (a) reconstructing conversational context with a fixed window ($k = 3$), (b) building a curated Ecuadorian-Spanish corpus, and (c) applying controlled synthetic augmentation with deduplication and cosine-similarity filtering, combined with cost-sensitive fine-tuning to improve minority-class recovery under a leakage-free evaluation protocol.

2.1. Digital Violence Dynamics and Linguistic Challenges

Digital violence is defined as behaviors of aggression, harassment, and discrimination amplified by virality and anonymity. In socio-occupational contexts, these dynamics are intrinsically related to moral harassment, or mobbing [2]. Recent initiatives such as EXIST have expanded the analysis toward multimodal content and the Learning with Disagreement paradigm, acknowledging that subjectivity is inherent to the task [16–18].

However, a critical computational challenge remains: indirect aggression (e.g., persistent degradation, sarcasm) requires contextual annotation that considers the dialogue structure, since analyzing isolated tweets is insufficient [19]. Dialectal variability adds further pragmatic complexity, as pretrained models suffer severe degradation in zero-shot transfer scenarios across Spanish variants [20]. To mitigate these issues, this study implements a context window approach ($k = 3$)—detailed in Section 4.1—which enables the model to capture both complex pragmatic structures and the local slang that acts as a crucial regional modifier [8].

2.2. Data Acquisition: The Post-API Paradigm

Corpus collection for social research has undergone a drastic shift known as the "APIcalypse" [15], characterized by the restriction or closure of academic access to the APIs of platforms such as X (formerly Twitter). This limitation has forced the scientific community to migrate toward advanced web scraping techniques.

Unlike traditional data mining, modern scraping must simulate human browsing (waiting times, scrolling, interface interaction) to bypass anti-bot mechanisms [21]. Methodologically, this technique is indispensable not only to obtain the text, but also to reconstruct the conversational thread. Theory suggests that, to detect indirect violence, it is critical to analyze the prior context (the N previous interactions), since an apparently neutral comment can become violent when identified as a response to a provocation or a harassment campaign.

2.3. Synthetic Data in NLP: Generation and Filtering

Given the scarcity of annotated corpora and the extreme class imbalance in social media, data augmentation is a vital strategy. However, indiscriminate generation can introduce noise. Therefore, this study adopts a three-phase protocol validated in recent literature to mitigate hallucinations and ensure quality:

1. Seed Selection: A Human-in-the-Loop approach is used, strictly relying on real examples manually validated (Gold Standard). This is justified in the literature because generative models tend to amplify input biases; consequently, using high-quality seeds is the only mechanism to anchor generation to the linguistic reality of the target dialect and avoid semantic degradation [22].

2. Controlled Generation (In-Context Learning): The Few-Shot Prompting is employed to leverage the ability of LLMs to recognize patterns at inference time. Specifically, previously validated instances are used as seed examples within structured prompts that instruct the model to preserve the communicative intent, semantic category, and linguistic characteristics of Ecuadorian Spanish while generating new variants. This approach enables the induction of local slang, morphology, and discourse patterns from a small number of demonstrative examples (*k*-shots) without modifying the model's weights [23], thereby mitigating the semantic degradation commonly observed in zero-shot scenarios. The generated samples are subsequently subjected to validation, deduplication, and semantic similarity filtering before being incorporated into the final corpus.
3. Filtering and Refinement: Synthetic generation often produces high redundancy rates. Based on evidence that language models memorize duplicated data, degrading their generalization ability [24], we implement exact deduplication and cosine-similarity filters. This ensures that synthetic data contributes real semantic variability to training, rather than merely adding volume that induces overfitting.

2.4. State-of-the-Art Architectures and Domain Adaptation

After assessing the availability of pretrained language models (PLMs), both multilingual and Spanish-specific, we analyzed several alternatives, including Transformer-based architectures and lighter statistical models. Among these, we examined the performance of probabilistic models such as LDA (Latent Dirichlet Allocation); however, its approach was found to be unable to fully capture the complex semantic context required to identify harassment situations in which word meaning varies by usage [12].

Similarly, we evaluated lightweight models from the BERT family, specifically DistilBERT, with the aim of optimizing computational resources. However, this architecture was discarded after showing critical difficulties in identifying instances for each class, yielding a limited accuracy between 34% and 45% [25]. This high dispersion in its results, compared to larger-scale models, confirms that detecting regional digital violence requires the full parameters of base-sized models to capture nuanced linguistic regularities. Based on these findings, we selected the following discriminative architectures due to their robustness and superior performance:

- mBERT (Multilingual): Multilingual BERT in its cased variant is built on the base-sized Transformer architecture and pretrained in a self-supervised manner on raw Wikipedia text in 104 languages [14,26]. To mitigate resource imbalance across languages, pretraining applies a sampling scheme that reduces the dominance of high-resource languages and increases exposure to low-resource languages. The model uses WordPiece tokenization with a shared vocabulary (110,000 subwords), is case-sensitive, and represents segment pairs as `[CLS]` Sentence A `[SEP]` Sentence B `[SEP]` (maximum length 512). Learning follows the standard objectives of Masked Language Modeling (MLM) and Next Sentence Prediction (NSP): in MLM, 15% of tokens are masked (80% `[MASK]`, 10% random replacement, and 10% unchanged), and in NSP the model predicts whether B is the continuation of A (50% real pairs and 50% random pairs) [14].
- BETO (Monolingual): Considered the gold standard for Spanish NLP, it corresponds to the "dccuchile/bert-base-spanish-wwm-cased" variant, pretrained exclusively on Spanish using Whole Word Masking. Comparative studies show that it captures the language's morphology, idioms, and syntactic regularities with higher fidelity, consistently outperforming mBERT in Spanish discourse-classification tasks on social media comments [13,27,28].

Fine-Tuning and Cost-Sensitive Learning

These models are not used directly; instead, they are adapted via supervised fine-tuning, adjusting the Transformer's weights to specialize general linguistic representations to the digital-violence taxonomy [8,14]. This phase is crucial for handling high-variation phenomena (spelling, abbreviations) typical of social media [29,30].

However, given the extreme imbalance in which critical classes are minority categories, standard training can fall into the "accuracy paradox" [9]. To counteract this, we incorporate a cost-sensitive learning approach using a weighted cross-entropy that assigns weights inversely proportional to class frequency. Additionally, to stabilize training and prevent overfitting under these conditions, we apply an optimized regularization strategy. This includes a weight decay range of 0.01–0.05, a dropout of 0.3, and label smoothing ($\alpha = 0.1$), thereby reducing overconfident predictions that could otherwise lead to gradient instability.

2.5. Related Work and Rationale for the Taxonomy

The selection of the four fundamental topics used in this study is not arbitrary; rather, it represents the consolidation of findings obtained in prior research under the CRISP-DM methodology [12,25]. This structure makes it possible to address the linguistic particularities and sociocultural nuances of the Ecuadorian context without incurring semantic redundancies.

In an initial study on mobbing as a cyber threat, eight general phases of workplace harassment were originally analyzed [12]. However, through a process of data refinement and perplexity analysis, it was determined that a structure of four specific clusters (managerial conflict, communication, influence on relationships, and emotional impact) achieved higher precision in capturing the essence of the phenomenon, removing noise detected through cosine similarity [12,25].

Subsequently, this architecture was validated in the analysis of Gender-Based Violence (GBV) in social media. By evaluating model coherence, it was empirically shown that a four-topic configuration could group information without overlaps between categories, facilitating a hierarchical classification of aggressiveness into three levels [25]:

- Level 0 (Non-Offensive): Texts that do not contain inappropriate or hostile content [25].
- Level 1 (Dark Humor): Ironic expressions, sarcasm, or aggressive jokes that normalize hostility without constituting direct threats [25]. This category poses a significant classification challenge because its pragmatic meaning often depends on the immediate conversational context, making isolated comments difficult to distinguish from neutral or benign interactions. To reduce this ambiguity, the proposed framework incorporates the context window ($k = 3$) described in Section 4.1, enabling the model to recover the preceding conversational cues required to correctly infer the speaker's intent.
- Level 2 (Intolerance): Discriminatory, dehumanizing, or stigmatizing language directed at groups or identities [25].
- Level 3 (Violence): Direct threats, incitement, or explicit descriptions of physical/sexual harm [25].

This hybrid methodology, combining LDA for topic delimitation and RoBERTa for contextual analysis, has achieved accuracies between 93% and 99% in identifying abusive behaviors, establishing the four topics as the most robust foundation for deploying content filters in regional scenarios [25].

2.6. Metrics and Evaluation Protocols

In content-moderation domains, evaluation must prioritize the ability to recover critical classes over overall accuracy, since the cost of a false negative in high-violence

categories is significantly higher [9,31]. Therefore, we implement a multidimensional evaluation protocol that includes the following indicators:

1. Global Performance Metrics:
   - Accuracy: Provides a measure of the model's overall correctness across all categories [32].
   - Weighted Precision, Recall, and F1-Score: Due to the inherent imbalance in social media data, weighted means are used to adjust each class's contribution according to its representativeness (support), preventing performance in the majority class (Non-Offensive) from masking deficiencies in critical classes [33,34].
2. Per-Class Classification Report: To assess the model's granularity within the four-class taxonomy, we individually analyze Precision, Recall, and F1-Score for each label [35]. This analysis is vital to identify ambiguity phenomena, such as those previously observed in the Dark Humor category, where distinguishing aggressive irony/sarcasm from neutral comments is often difficult [8,36]. Support is included to validate the statistical significance of the results in each category.
3. Training and Convergence Metrics: We monitor Training Loss and Validation Loss to supervise the Transformer's learning process [14]. A controlled gap between the two losses is indicative of good generalization capability and helps prevent overfitting [37].

Finally, validation is supported by qualitative analysis using confusion matrices, which makes it possible to visualize intersections among classes and tune the model to minimize operational risks [38]. Based on theoretical recommendations on balanced partitions [39], we implemented a two-stage stratified scheme, resulting in a final distribution of 70% for training, 15% for validation, and 15% for testing. Crucially, as detailed in Section 4.3.1, this split was performed prior to any synthetic data augmentation. Consequently, synthetic instances were generated exclusively from the training partition, while the validation and test sets remained composed solely of original data. This protocol prevents data leakage and ensures that the final evaluation is conducted on completely unseen, non-synthetic samples.

2.7. Research Questions

Based on the described context, this work is guided by the following research questions:

- RQ1: Which techniques enable appropriate selection and curation of data in Latin American Spanish?
- RQ2: Which methodological approach can be used for model selection in Spanish?

3. Methodology

3.1. Studied Problem (Task Definition)

We study the automatic detection of digital violence in Ecuadorian Spanish as a multi-class inappropriate-content classification problem under strong dialectal variation and extreme class imbalance. Given a target social-media comment t_0 and its conversational context window ($k = 3$ preceding interactions, t_{-1}, t_{-2}, t_{-3}), the goal is to predict one of four labels: Non-Offensive, Dark Humor, Intolerance, or Violence. The main technical difficulties are (i) implicit aggression (sarcasm, irony, regional idioms) that requires context, and (ii) the low prevalence of critical classes in naturally occurring streams, which makes naive training and accuracy-only evaluation unreliable.

3.2. Methodological Framework

This study adopts the standard CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) as the guiding thread to organize the technical development. Figure 2 summarizes the adapted workflow and its four phases. Given the unstructured

nature of the data and the scarcity of resources in the target dialect, the workflow is adapted into four interconnected operational phases:

- Problem and Data Understanding: Definition of the contextual digital-violence classification task and exploratory analysis of class distribution (severe imbalance) in the Ecuadorian digital ecosystem.
- Data Preparation and Engineering: Design of a hybrid (Human–AI) pipeline that goes beyond traditional cleaning. It includes post-API corpus acquisition, manual label validation after the failure of generic classifiers, and the implementation of a synthetic data generation and filtering protocol to correct class imbalance.
- Modeling: Selection of a monolingual Transformer architecture optimized for Spanish to execute supervised fine-tuning. This phase involves a critical performance comparison between training with the original imbalanced data versus our curated hybrid corpus, specifically testing the hypothesis that semantic deduplication and filtering are essential for effective domain adaptation in regional violence detection.
- Evaluation: Based on theoretical recommendations on balanced partitions [39], a two-stage stratified scheme was implemented, resulting in a final distribution of 70% for training, 15% for validation, and 15% for testing, performed before any synthetic augmentation (Section 4). Specifically: (i) a first stratified split with `test_size` = 0.30 reserves 70% for training; and (ii) the remaining 30% is stratified-split with `test_size` = 0.50, yielding two equal halves (15% validation, 15% test). Validation is used for model selection (early stopping), and the test set is reserved for the final evaluation, prioritizing per-class metrics (precision, recall, and F1) and error analysis via confusion matrices.

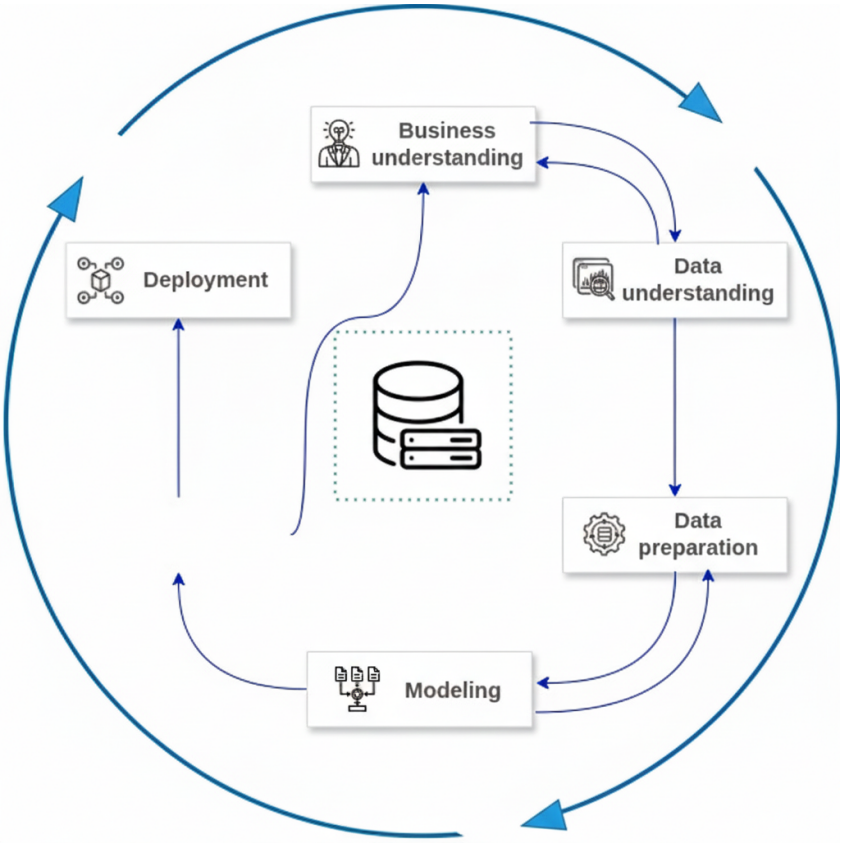


Figure 2. CRISP-DM guided workflow adapted for digital-violence detection with synthetic data augmentation.

3.3. Detailed Workflow (Step-by-Step)

To address the request for a more detailed methodology description, we summarize the concrete workflow executed in this study as an ordered sequence of steps (inputs, operations, and outputs). Each step is described in detail in Sections 4 and 5.

1. Data acquisition and thread reconstruction: collect posts and their conversational context under post-API restrictions, and build context windows with $k = 3$ preceding interactions (Section 4).
2. Initial labeling and validation: perform a zero-shot labeling attempt and then apply manual validation to establish a reliable ground truth and quantify model failure under dialectal variation.
3. Train/validation/test split (before augmentation): split the validated corpus into 70%/15%/15% using stratification, ensuring that any synthetic data derived from seeds remains within the training partition (leakage prevention).
4. Synthetic data generation (minority classes only): generate class-conditioned candidates from validated seeds using few-shot prompting (Section 4).
5. Quality control and filtering: remove exact duplicates and eliminate semantically redundant candidates via cosine-similarity filtering, keeping only samples that contribute semantic variability.
6. Dataset assembly: construct the final training corpus by leveling the minority classes to a target operational threshold while keeping the validation and test partitions purely original.
7. Model selection and training scenarios: compare BETO and mBERT under a controlled baseline and train BETO under multiple imbalance-handling strategies (original, undersampling, oversampling, and cost-sensitive learning) (Section 5).
8. Evaluation and analysis: report global and per-class metrics, confusion matrices, and threshold-robust curves (ROC and Precision–Recall), emphasizing minority-class recovery in imbalanced settings (Section 6).

The structure of the paper reflects this methodological flow: the technical details of acquisition, the analysis of LLM failure, and synthetic data engineering are described in depth in Section 4. The model architecture, fine-tuning hyperparameters, and training strategies are presented in Section 5. Finally, Section 6 reports the empirical evidence of improved detection enabled by the proposed workflow.

4. Development: Data Acquisition and Processing

The quality of violence-detection models depends directly on the representativeness of the training corpus. Figure 3 provides a visual roadmap of the end-to-end data engineering pipeline used in this study, from post-API acquisition and conversational-context reconstruction to human validation and iterative class leveling via curated synthetic generation. To provide a clear methodological overview, this section is structured as follows: (i) acquisition and conversational-context reconstruction, (ii) labeling analysis and human validation, (iii) split protocol to prevent leakage, (iv) synthetic generation and filtering, and (v) final dataset composition.

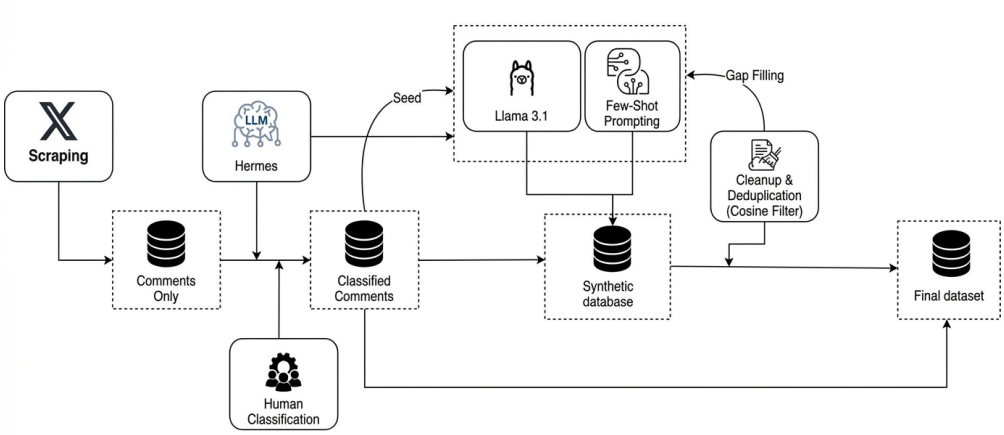


Figure 3. Data Engineering Pipeline with Iterative Class Leveling. The flow integrates extraction (scraping), preliminary classification (Hermes), and human validation. The lower branch details the cyclic process for synthetic data: generation with LLaMA 3.1, strict cleaning (deduplication + cosine similarity), and a leveling loop (gap filling) that reinjects seeds until reaching the operational threshold of 25,000 records per class.

4.1. Acquisition and Contextualization Pipeline

The acquisition strategy was designed to operate under restrictions on access to official APIs (APIcalypse) [15]. We implemented a modular scraper based on Selenium and ChromeDriver [21], designed to collect not only isolated posts, but complete conversational threads.

A critical component was the definition of a context window ($k = 3$). For each target comment (t_0), the system stored the three immediately preceding interventions (t_{-1}, t_{-2}, t_{-3}), enabling the partial reconstruction of the conversational thread. This design directly addresses the research gap identified in Section 2.1, namely the inability of message-level approaches to capture context-dependent pragmatic information. By incorporating the immediate conversational history, the model is better equipped to interpret sarcasm, irony, dark humor, and implicit aggression, reducing semantic ambiguity and improving the inference of the speaker's communicative intent.

- Source: High-visibility and highly polarized profiles in Ecuador (politics, entertainment, crime news).
- Volume: 267,315 raw records were obtained, consolidated into 238,007 unique records after normalization.

4.2. Labeling Analysis: Zero-Shot Evaluation and Human Validation

Initial labeling was delegated to the Hermes 2 Pro and Hermes 3 models (based on LLaMA 3.1) under a zero-shot scheme. Although the literature suggests that LLMs can compete with human annotators [40], our experiment revealed severe performance degradation in the Ecuadorian dialect domain.

Manual validation (ground truth) of the entire corpus exposed a critical discrepancy. As illustrated in Figure 4, the model exhibited a massive rate of false positives in the "Violence" class:

- Automatic Prediction: 18,531 records classified as violence.
- Validated Reality: Only 347 records were genuine violence.
- Error Rate: 97.5% of alerts were cultural-interpretation errors (use of coarse slang in non-violent contexts).

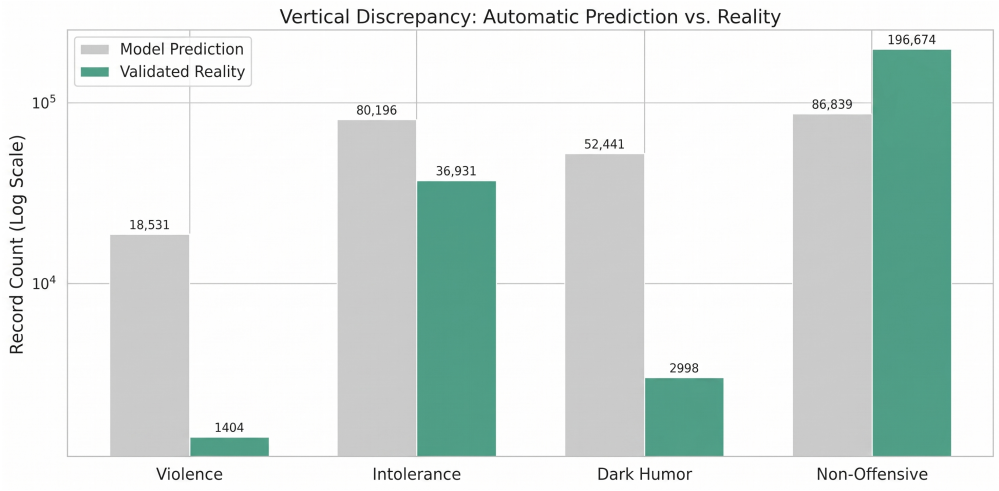


Figure 4. Discrepancy between Automatic Prediction and Reality. Note the logarithmic scale: the model overestimated violence by a factor of 50×, forcing the automatic labeling to be discarded.

This phase confirmed an extreme class imbalance (approximately 200:1 between neutrality and violence), making it unfeasible to train a robust discriminative classifier without intervention.

The manual validation of the entire corpus was conducted by three native Ecuadorian co-authors with backgrounds in systems engineering and computational linguistics. The annotation process followed structured guidelines based on the four-level digital violence taxonomy defined in Section 2.5. To guarantee the absolute integrity of the Gold Standard, we established a collaborative, consensus-driven adjudication protocol rather than independent statistical ratings subject to marginal agreement. Under this scheme, any comment presenting ambiguity, sarcasm, or borderline categorization was collectively reviewed and debated by the experts until a 100% unanimous consensus was reached. This collaborative expert validation ensures a single, highly accurate, and noise-free ground truth, which is critical when dealing with implicit aggression and complex local idioms.

4.3. Synthetic Data Engineering to Correct Class Imbalance

To level minority classes (Violence and Dark Humor) without introducing artifacts that invalidate evaluation, a strict generation and curation protocol was executed.

4.3.1. Split Protocol (Preventing Data Leakage)

To ensure methodological integrity, the validated seed corpus was partitioned into training (70%), validation (15%), and test (15%) sets prior to any synthetic data generation. This distribution provides a balanced allocation of data for model learning, hyperparameter tuning, and unbiased performance evaluation, following established recommendations for supervised learning workflows [41]. Furthermore, performing the partition before data augmentation is essential to prevent "data leakage", as all synthetic instances derived from a given seed remain within the same subset [42]. Consequently, the validation and test sets contain only original, unseen samples, ensuring that the model is never evaluated on synthetic variations of examples encountered during training and that the reported performance accurately reflects its generalization capability.

4.3.2. Generation and Filtering Strategy

A hybrid scheme was used, applying few-shot prompting on LLaMA 3.1 [22], followed by an aggressive cleaning phase to mitigate redundancy inherent to generative models.

1. Seeding: Use of the 1404 real violence records and 2998 real dark-humor records as a base.
2. Mass Generation: Production of class-conditioned variations.
3. Cleaning (Deduplication + Cosine Similarity): To mitigate the redundancy inherent to LLM-based generation, we applied a two-stage cleaning procedure combining exact deduplication and semantic similarity filtering. After removing identical records, sentence embeddings were generated using the `hiiamsid/sentence_similarity_spanish_es` model, and a cosine-similarity threshold of $\tau = 0.85$ was applied to identify semantically redundant samples. As shown in Figure 5, approximately 45% of the generated records (around 17,000 instances) were discarded because, despite exhibiting superficial lexical differences, they conveyed essentially the same semantic content as previously existing samples. Retaining these instances would have artificially increased the size of the corpus without contributing new information, reducing linguistic diversity and increasing the risk of overfitting. Consequently, the resulting hybrid corpus preserves genuine semantic variability, providing empirical support for the semantic richness introduced through the synthetic data generation process.
    - Violence: From 16,825 generated $\rightarrow$ 9154 unique.
    - Dark Humor: From 20,821 generated $\rightarrow$ 11,540 unique.

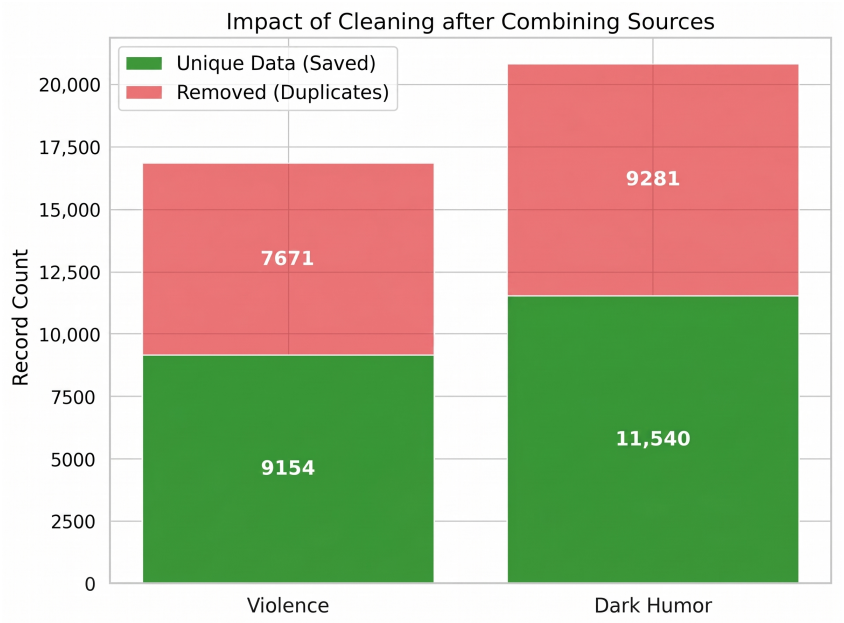


Figure 5. Impact of Quality Filtering. Thousands of redundant records (red bar) were removed to prevent the model from memorizing repetitive patterns.

To qualitatively verify that the synthetic samples generated by LLaMA 3.1 preserved the semantic and pragmatic nuances of Ecuadorian Spanish, we performed a manual qualitative inspection on a randomized sample of 100 generated sentences from the minority classes (Violence and Dark Humor). The analysis confirmed that the few-shot prompting strategy successfully guided the model to retain natural colloquial structures and localized slang without introducing artificial syntax or grammatical translation artifacts. This dialectal naturalness, combined with the sentence embedding filter (`hiiamsid/sentence_similarity_spanish_es` with $\tau = 0.85$) that discarded 45% of redundant generations (Figure 5), ensured that the final augmented dataset introduced genuine semantic and lexical diversity rather than repetitive, overfitted patterns.

4.3.3. Final Dataset Composition (Gap Filling)

Finally, controlled generation was applied to close the gap until reaching the operational threshold of 25,000 instances for the minority classes (Violence and Dark Humor). Unlike traditional approaches that aggressively undersample the majority class, this study preserves the entirety of the Non-Offensive class (196,674 records). This design choice avoids the loss of valuable linguistic information and keeps the learning scenario closer to real social-media conditions, where harmful content is low-prevalence but high-impact.

The resulting training corpus (Figure 6) presents a hybrid structure:

- Augmented critical classes: Violence and Dark Humor are raised to 25 k by injecting 29,306 curated synthetic examples.
- Purely real classes: Intolerance (36k) and Non-Offensive (196 k) remain entirely real to stabilize training and reduce false positives.

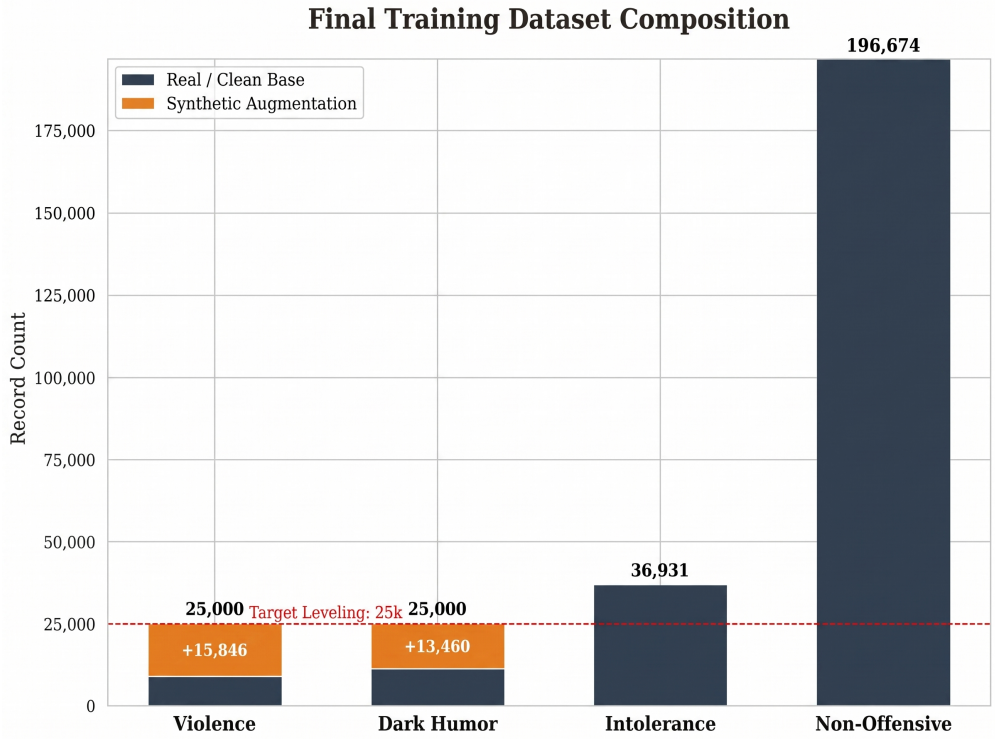


Figure 6. Final dataset distribution. Selective correction through curated synthetic data (orange) is observed against the massive Non-Offensive class (blue), configuring a realistic learning scenario.

5. Development: Model Selection and Modeling

5.1. Experimental Setup (Environment and Tools)

To ensure reproducibility, we report the environment and tools used in the experiments. The implementation is based on Python and standard NLP/ML libraries for Transformer fine-tuning (e.g., Hugging Face Transformers and PyTorch), and the evaluation uses common scientific-computing tools (e.g., scikit-learn for metrics and plotting). Data acquisition is performed via Selenium-based browsing automation (Section 4).

- Hardware: GPU: T4, RAM: [High-RAM].
- Software: Python [3.12], PyTorch [2.5.1], Transformers [4.46.1], scikit-learn [1.5.2].

5.2. Model Selection

Model selection was approached as an empirical decision guided by performance in Spanish and generalization capacity under class imbalance. We compared the two pre-

trained Transformer architectures described in Section 2: the monolingual BETO model and the multilingual alternative mBERT.

The comparison was carried out under controlled conditions using a balanced reference scenario (undersampling) to isolate the architectural effect. Based on the results reported in Section 6, BETO exhibited better F1-score and accuracy; therefore, it was selected as the final model for the remaining experiments.

5.3. Fine-Tuning with Mixed Data

The selected model was retrained via supervised fine-tuning on the mixed corpus, where curated synthetic data is used to level minority classes in the training set (Section 4). To enhance readability, this section is structured as follows: (i) preprocessing, (ii) training scenarios, (iii) parameter selection and final configuration, and (iv) evaluation protocol.

5.3.1. Preprocessing

Texts were tokenized with the model's associated tokenizer, normalizing the input by removing noise (multiple spaces, non-printable characters) and truncating sequences that exceed the maximum length, while preserving the target message and its extended context.

5.3.2. Training Scenarios

Four scenarios were evaluated: (1) the original imbalanced dataset, (2) balanced via undersampling, (3) with synthetic oversampling, and (4) cost-sensitive training via weighted cross-entropy. The cost-sensitive approach was prioritized because it penalizes errors in minority classes without distorting the real distribution.

5.3.3. Hyperparameter Selection and Final Configuration

Hyperparameters were chosen following standard recommendations for BERT-style fine-tuning and then validated on the held-out validation set (15%). When multiple configurations were compared, the selection criterion prioritized (i) validation loss and (ii) macro-averaged F1 to avoid majority-class dominance.

Fine-tuning was performed with a learning rate of 2×10^{-5}, batch size 16, maximum length 128 tokens, and weight decay 0.05. We used a Weighted Trainer integrating `CrossEntropy` loss with smoothed class weights (power 0.5) and label smoothing ($\alpha = 0.1$). Dropout was set to 0.3. Early stopping (patience 2 epochs) selected the best checkpoint based on validation loss.

5.3.4. Evaluation Protocol

Evaluation focused on per-class metrics (precision, recall, and F1) and confusion matrices, since overall accuracy is poorly informative in imbalanced scenarios. The primary success criterion is the recovery of digital-violence instances (reduction of false negatives), even at the cost of increased manual review due to false positives.

6. Results

Quantitative analysis reveals substantive differences depending on the data-handling strategy and confirms that, in Latin American contexts, the combination of extreme imbalance and linguistic variation can degrade the recovery of digital violence when training on raw data.

We evaluate performance using accuracy; weighted precision, recall, and F1; macro F1; and per-class precision/recall/F1 from the classification report. For error analysis, we include confusion matrices, and to assess robustness across decision thresholds, we report ROC and Precision–Recall curves.

First, the architecture comparison under controlled conditions (undersampled dataset) shows that the monolingual BETO model captures Spanish semantic regularities more faithfully. BETO achieved an accuracy of 79.43% and an F1-score of 0.7948, outperforming Multilingual BERT (mBERT) [14], which obtained an accuracy of 77.63% and an F1-score of 0.7763. This difference, although moderate in absolute value, is operationally relevant because this is a task sensitive to context and local slang (Figure 7).

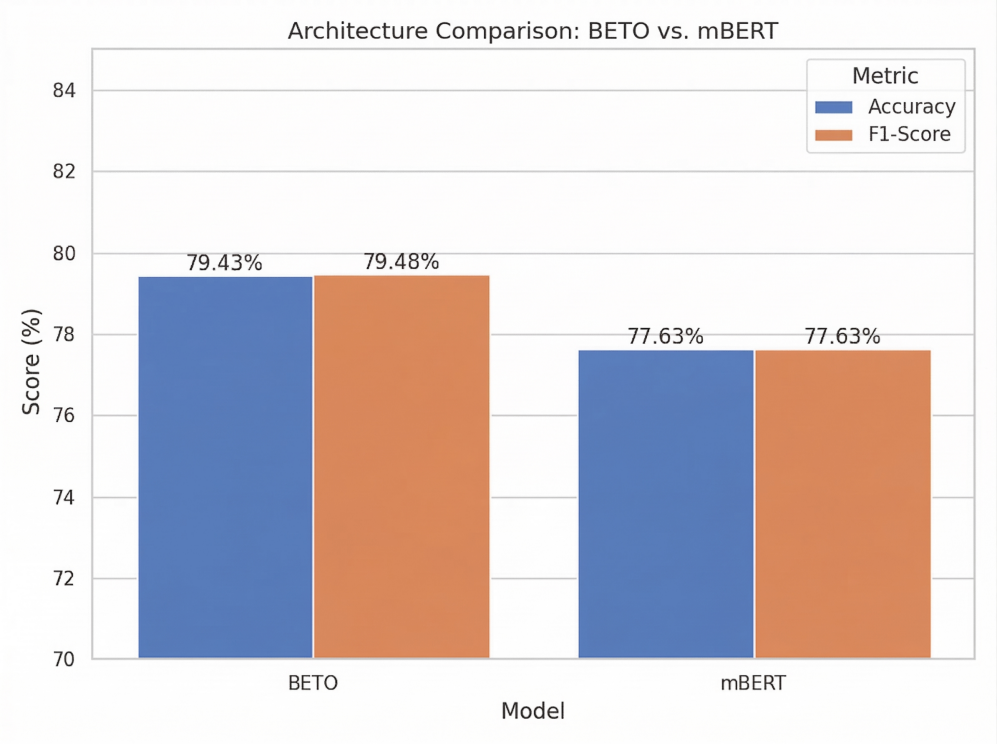


Figure 7. Performance comparison between architectures: BETO (79.43% accuracy) outperforms mBERT (77.63%) on the test set.

Second, training with the original imbalanced dataset confirms the "Accuracy Paradox": although overall accuracy was high (93.90% with BETO), the recovery of minority classes was poor, with a Recall of 18.09% for the Intolerance class. This pattern implies a high false-negative rate in critical categories and shows that accuracy alone is not an adequate indicator of safety/efficacy in content moderation (Figure 8).

While synthetic oversampling yielded seemingly perfect metrics (>99.6%), this behavior was identified as overfitting induced by data redundancy and leakage, creating an artificial separation in validation that does not generalize to real data (Figure 9). This 'artificial perfection' was decisive in discarding full oversampling as a viable strategy, as it represents a symptom of evaluation contamination that could lead to catastrophic failures in real-world deployments. To mitigate these risks, we transitioned toward the E4 scenario. Instead of unconstrained oversampling, this strategy implements a balanced hybrid approach: it combines regularized data augmentation with cost-sensitive learning via a Weighted Trainer. Consequently, the burden of correcting the class imbalance is shifted from mass artificial data generation to strategic algorithmic penalization.

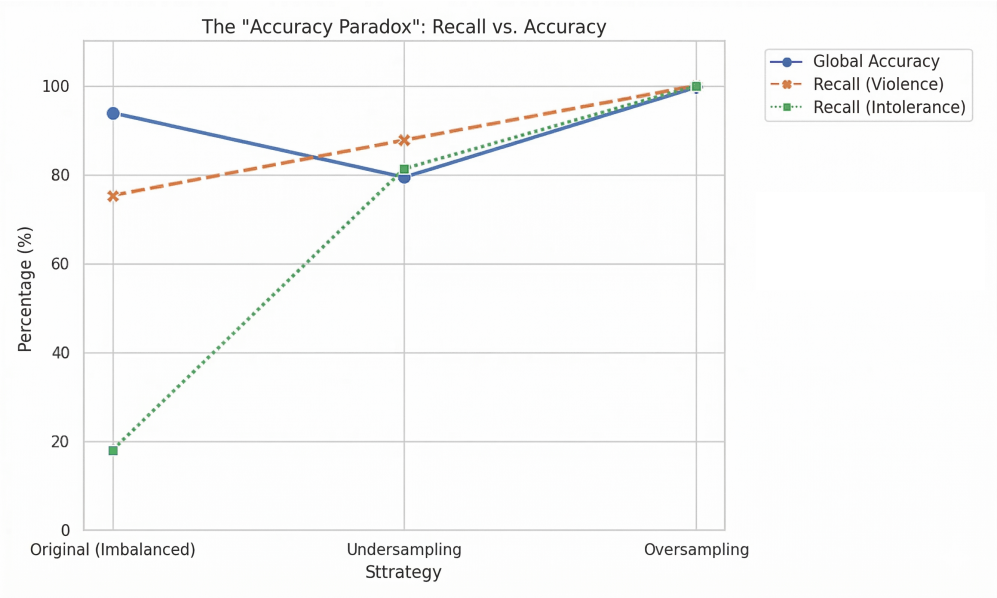


Figure 8. The "Accuracy Paradox": in the original dataset, high accuracy masks poor recall in minority classes.

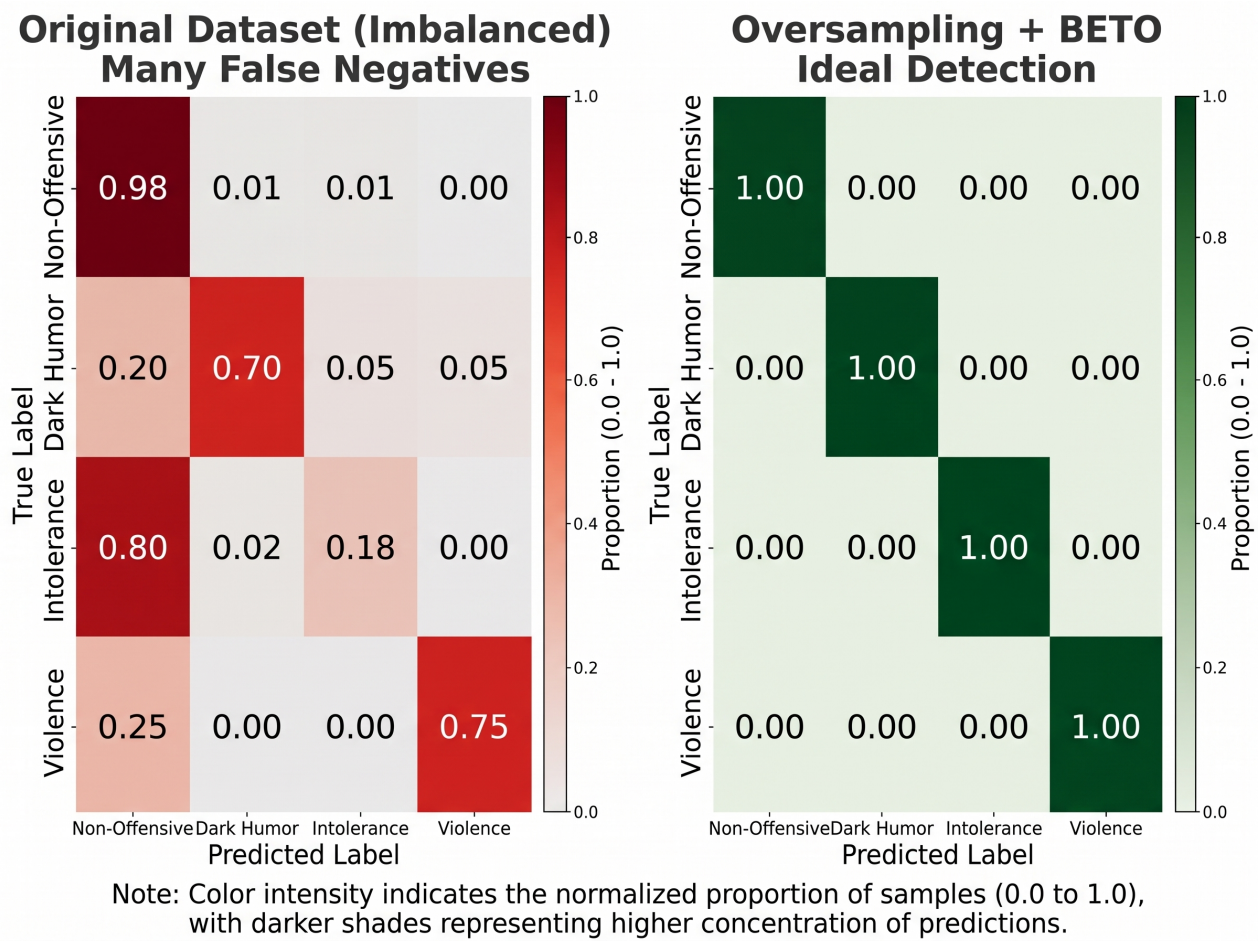


Figure 9. Impact analysis of balancing. Left: original dataset with high false negatives (inability to detect threats). Right: oversampling-induced overfitting, showing an artificial "perfection" that indicates lack of generalization.

This result (Figure 9) was decisive in fully eliminating oversampling as a training strategy for the final prototype: the apparent "perfection" of the matrix under oversampling is not evidence of predictive capability, but rather a symptom of contamination/redundancy that invalidates evaluation and can induce unsafe deployments. Consequently, the modeling progression shifted toward an approach that preserves the distribution and evaluation protocol, replacing resampling with cost-sensitive learning and regularization to improve the recovery of critical classes without compromising generalization.

In contrast, Scenario E4 (BETO-CSR+D) (BETO with semantic deduplication, cost-sensitive learning, and regularization) avoided the dichotomy between information loss (undersampling) and memorization (oversampling). On the test set, the model achieved an overall accuracy of 94.39%, with weighted F1-score of 0.9429 and Macro F1 of 0.9022. At the class level, F1-scores were: Non-Offensive (0.9669), Violence (0.9028), Intolerance (0.8801), and Dark Humor (0.8590). Overall, the results reinforce that the value of hybrid data depends on its curation (including semantic deduplication) and on an evaluation protocol free of contamination.

Figure 10 presents the confusion matrix of the final model (Scenario E4). Unlike the baseline model, an effective redistribution of errors is observed: the detection capability for critical classes increases without collapsing the precision of the majority class, consistent with cost-sensitive training and regularization.

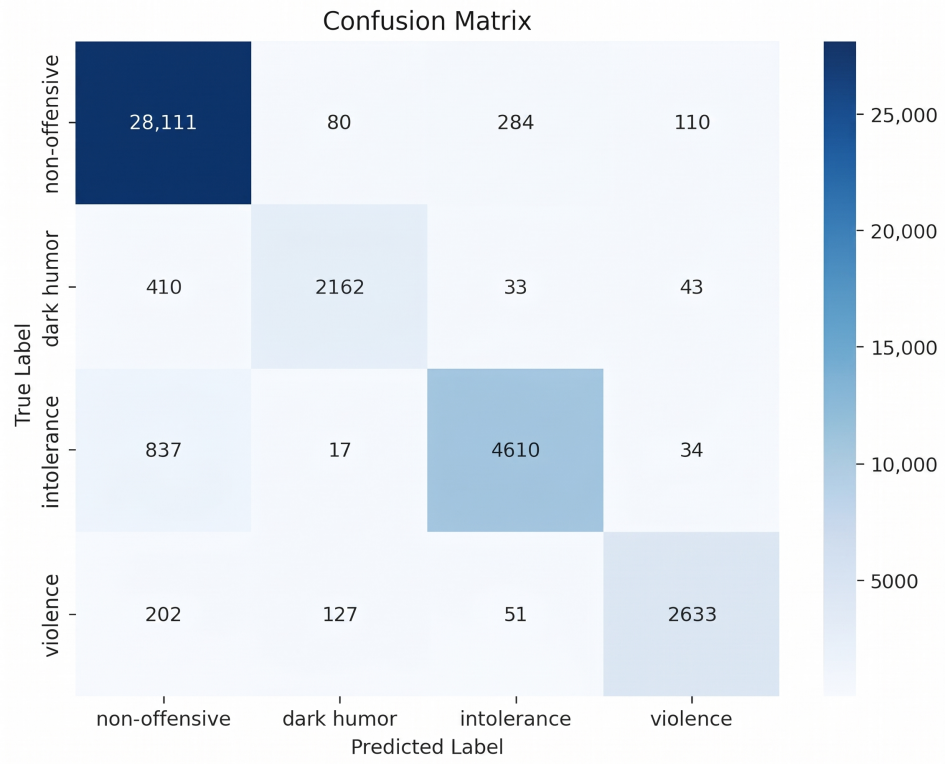


Figure 10. Confusion matrix of the final model (Scenario E4). An effective balance is observed between detecting minority classes and preserving the majority class.

To validate classification robustness beyond a fixed threshold, ROC curves (Figure 11) and Precision–Recall curves (Figure 12) were analyzed. ROC curves indicate high separability among classes. Precision–Recall curves, in turn, show that the model maintains acceptable precision even at high recovery levels, a balance that is difficult to achieve in highly imbalanced scenarios.

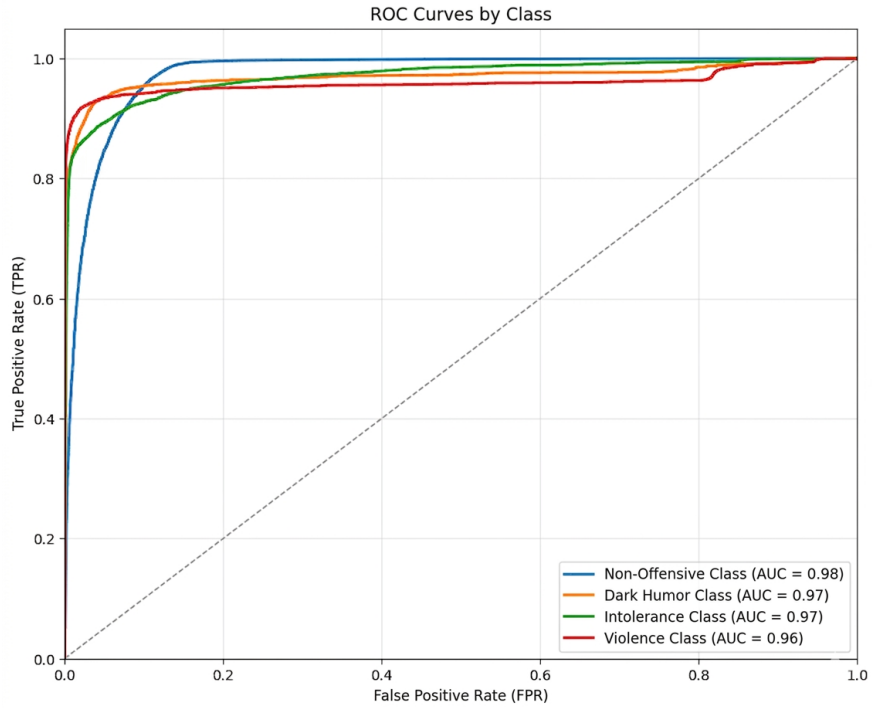


Figure 11. ROC curves of the final model (Scenario E4).

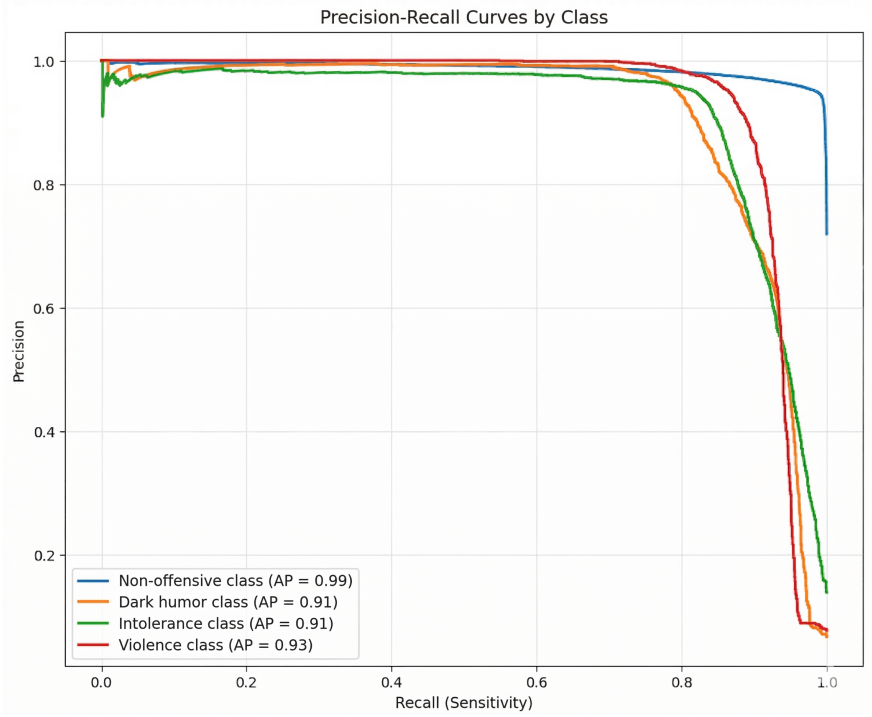


Figure 12. Per-class Precision–Recall curves (final model, Scenario E4).

7. Discussion

7.1. Data Selection and Curation in Latin American Spanish (RQ1)

One of the most revealing findings of this study is the confirmation of the "Accuracy Paradox" in cybersecurity datasets. The model trained with original data achieved an accuracy of 93.90%, a figure that superficially appears excellent, yet its ability to detect *Intolerance* was almost null (Recall: 18.09%). This behavior shows that proper data selection in Latin American Spanish contexts must explicitly address class imbalance and preserve the validity of the evaluation protocol.

When contrasting traditional resampling techniques, we observed that while undersampling improves sensitivity, it sacrifices too much contextual information, and naive oversampling leads to severe overfitting. In particular, experimentation with oversampling revealed a critical methodological risk: data leakage. By obtaining near-perfect metrics (>99%) in validation, it became evident that the model was not learning to generalize, but rather to memorize synthetic patterns repeated across the training and test sets. This underscores that LLM-based data augmentation (e.g., with LLaMA 3.1) must be applied under rigorous split protocols before synthetic generation. Blind reliance on synthetic data without quality controls can lead to a false sense of security, deploying models that fail catastrophically on new real-world data.

In particular, the confusion matrix for the oversampling scenario (Figure 9) exhibits an "ideal separation" pattern that, in this context, is interpreted as an evaluation artifact rather than a genuine modeling gain. For this reason, oversampling was used exclusively as a diagnostic baseline and was fully discarded in the construction of the final prototype.

Methodologically, the progression was consolidated as follows: (i) a baseline was established with the original dataset to reveal the accuracy paradox; (ii) undersampling was evaluated to isolate the architectural effect and compare BETO vs. mBERT; (iii) Scenario E3 was defined as cost-sensitive and regularized training (dropout, label smoothing, and smoothed class weights) to improve recovery without distorting the distribution; and (iv) Scenario E4 was defined by additionally incorporating semantic deduplication to reduce redundancy and improve generalization.

Scenario E4 addresses this dichotomy by using a specialized architecture (BETO) together with cost-sensitive learning and explicit regularization, incorporating label smoothing and smoothed class weights to reduce overconfident predictions and stabilize training under imbalance. In evaluation, the model achieved an accuracy of 94.39%, with a 0.9429 weighted F1-score and a 0.9022 Macro F1. At the class level, Violence achieved an F1-score of 0.9028, with Precision of 93.37% and Recall of 87.39%. Operationally, although recall is slightly lower than in previous configurations, the high precision implies a more reliable system with a lower false-alarm rate, which is preferable in real environments where the cost of human review and alert fatigue are critical factors.

Scenario E3 addresses this dichotomy by using a specialized architecture (BETO) together with cost-sensitive learning and explicit regularization (dropout), label smoothing, and smoothed class weights, which reduces overconfident predictions and stabilizes training under imbalance. In evaluation, the model achieved an accuracy of 92.68% and a 0.9266 weighted F1-score, with 92.66% recall for Violence and 81.32% for Intolerance. Operationally, the improvement is not limited to an aggregate metric, but translates into a substantial recovery of critical threats, mitigating the previously observed paradox.

Answer to RQ1: Appropriate selection and curation of data in Latin American Spanish requires (i) strict split protocols to prevent data leakage, especially before generating synthetic data; (ii) quality controls (normalization, filtering, and deduplication, including semantic deduplication) to reduce redundancy and generator biases; and (iii) training strategies that mitigate imbalance without distorting evaluation. In this work, naive oversampling was discarded because it induces evaluation artifacts, and the robust solution was cost-sensitive and regularized training (Scenario E4), which improved overall performance while maintaining a favorable balance between sensitivity and reliability.

7.2. Model Selection in Spanish (RQ2)

For model selection in Spanish, an empirical approach of controlled comparison was used among pretrained architectures, evaluated under the same protocol and with equivalent metrics. Within this framework, results confirm that linguistic specialization

is a determining factor in detecting verbal violence. The BETO model [13] consistently outperformed mBERT on all key metrics (F1-score: 0.7948 vs. 0.7763 in balanced settings). Although numerically moderate, this difference is qualitatively critical. Error analysis suggests that mBERT tends to fail on expressions that depend on local Ecuadorian slang or complex grammatical constructions (sarcasm), whereas BETO, having been pretrained exclusively in Spanish, has a richer vector representation of the language's morphology and semantics. Consequently, for culturally context-sensitive NLP tasks such as mobbing, monolingual models provide a solid reference when the target domain is Spanish.

Answer to RQ2: A selection procedure based on controlled comparison of pretrained models for Spanish (same splits, same evaluation protocol, and per-class metrics) is recommended. In this study, such comparison showed consistent advantages of BETO over mBERT; therefore, BETO was adopted as the final architecture.

It is important to emphasize that the core contribution of this work is not to benchmark the expanding ecosystem of Spanish Pretrained Language Models (PLMs), but to propose and validate a structured, CRISP-DM-guided data engineering pipeline for mitigating extreme class imbalance in regional dialects. Within this scope, BETO and mBERT represent the two fundamental paradigms in state-of-the-art Spanish NLP. The fact that the monolingual model (BETO) achieved an overall Macro F1 of 0.9022 (Table 1) under our curated hybrid scenario (E4) demonstrates the effectiveness of our data-centric pipeline. Exploring newer PLMs (such as Spanish RoBERTa variants) remains a valuable direction for future work but falls outside the immediate scope of this methodological framework.

Table 1. Overall metrics and per-class F1 (Scenario E4, test set).

Metric	Value
Accuracy	94.39%
Weighted F1	0.9429
Macro F1	0.9022
F1 Non-Offensive	0.9669
F1 Violence	0.9028
F1 Intolerance	0.8801
F1 Dark Humor	0.8590

7.3. Critical Analysis and Limitations

Although the inclusion of a context window ($k = 3$) proved fundamental to capturing indirect aggression, the integration mechanism in BETO presents inherent structural limitations when processing multi-part dialogues. Flattening conversational history into a single input sequence risks diluting the temporal hierarchy and critical distinctions between speaking turns. This architectural constraint suggests that the model's high performance may rely heavily on detecting rich lexical co-occurrences within the expanded context, rather than fully parsing the complex sequential pragmatics of aggression. Rather than a drawback, this limitation opens a critical avenue for architectural evolution. Future iterations can leverage this baseline by integrating hierarchical attention networks or dedicated speaker embeddings to explicitly model dialogue structures, transforming current sequence-structuring limitations into a framework for robust, multi-party conversational tracking.

On the other hand, the decision to discard oversampling in favor of a cost-sensitive approach proved methodologically superior to avoid memorization-driven overfitting. Nevertheless, performance on Violence should be interpreted with caution: although the model achieves high precision (93.37%) and a high F1 (0.9028), recall (87.39%) indicates that false negatives persist. Consequently, future iterations should complement evaluation

with cross-validation on other Spanish variants and out-of-domain analysis to verify that the model is not exploiting topic- or style-specific regularities of the hybrid dataset.

7.4. Practical Implications

For content moderators and social platforms, these results suggest that there is no "silver bullet." The optimal solution is not simply "more data" (oversampling), but rather "better data" (balancing and curation).

Operational limitations and external validity. These findings should be interpreted in light of four practical considerations: (i) the experimental protocol uses a stratified Train/Validation/Test split (70/15/15) with a fixed seed, implemented via a double split: first `test_size`=0.30 to separate 70% for training and, on the remaining 30%, `test_size`=0.50 to obtain two equal halves (15% validation, 15% test). The Validation set is used for model selection (early stopping) and Test is reserved for final evaluation, reducing the risk of bias from decisions guided by the test set; (ii) results are reported for a single seed (`random_state=42`). While multi-seed sensitivity analysis is a standard requirement for industrial deployment, the statistical reliability of our reported Macro F1 of 0.9022 is mathematically supported by our strict split protocol. Since the 70/15/15 stratified split was executed strictly prior to any synthetic augmentation, our test partition is 100% real and remained completely isolated from synthetic patterns. Evaluating on a purely authentic, unseen distribution provides a highly demanding and unbiased metric, ensuring that the reported performance gains are robust against the random initialization variations typically amplified by data leakage; (iii) although Scenario E4 explicitly included semantic deduplication (embedding model and threshold $\tau = 0.90$), performance may be sensitive to the choice of threshold and the redundancy distribution of the domain; thus, a sensitivity analysis and replication under multiple configurations remains future work; and (iv) evidence of generalization is limited to the Ecuadorian domain, so cross-validation with other Spanish variants (e.g., Mexico/Chile/Spain) is required to support broad claims of robustness.

In this framework, the proposed BETO-based architecture with cost-sensitive and regularized training offers a viable trade-off for implementation in early-warning systems. In particular, the Violence profile in Scenario E4 (precision 93.37% and F1 0.9028) suggests a reduction in false alarms while maintaining high recovery, improving the system's operational acceptability. For a safe deployment, it is recommended to:

(a) keep real Validation and Test sets strictly separated,
(b) report stability across multiple seeds and confidence intervals, and
(c) validate the model in out-of-domain scenarios and/or other dialectal variants.

8. Conclusions

This research addressed the problem of automatic detection of workplace mobbing and digital violence in Spanish, a field characterized by resource scarcity and linguistic complexity. Through a rigorous experimental design grounded in the CRISP-DM methodology, deep learning architectures and data-handling strategies were evaluated to answer the stated research questions.

We conclude that the monolingual BETO [13] architecture is superior to multilingual alternatives (mBERT) for this specific task, achieving better capture of semantic and cultural nuances of harassment in Spanish. Likewise, class imbalance was shown to be the largest technical obstacle in content moderation. Although traditional strategies such as undersampling and oversampling were evaluated as comparative baselines, they presented inherent limitations in terms of information loss or overfitting.

In contrast, the proposed final configuration (Scenario E4: cost-sensitive learning with regularization and semantic deduplication) emerged as the most robust approach, achieving an accuracy of 94.39%, with a 0.9429 weighted F1-score and a 0.9022 Macro F1. At the class level, F1-scores show consistent performance across all categories (Non-Offensive: 0.9669; Violence: 0.9028; Intolerance: 0.8801; Dark Humor: 0.8590), reinforcing that evaluation should focus on per-class metrics rather than only overall accuracy.

Finally, this study warns about the dangers of unregulated use of synthetic data. While LLMs offer enormous potential for data augmentation, incorrect application (naive oversampling) leads to severe overfitting that invalidates the model's practical utility. Future research lines should explore hybrid approaches that combine transfer learning with more advanced few-shot learning techniques and the incorporation of multimodal analysis (text and image) for comprehensive cyberbullying detection.

Author Contributions: Conceptualization, P.X.Z.R.; methodology, P.X.Z.R.; software, C.E.A.V.; validation, C.E.A.V.; formal analysis, M.P.S.A.; investigation, J.S.I.M.; resources, A.D.O.C. and A.E.P.M.; data curation, A.D.O.C. and A.E.P.M.; writing—original draft preparation, J.S.I.M.; writing—review and editing, M.P.S.A.; visualization, J.S.L.E.; supervision, P.X.Z.R.; project administration, P.X.Z.R. All authors have read and agreed to the published version of the manuscript.

Funding: The article processing charge (APC) for this publication is financially supported by the Vicerectorado de Investigación, Innovación y Vinculación of the Escuela Politécnica Nacional (EPN), Ecuador, through its institutional publication support program. No specific grant number is associated with this funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy limitations concerning the use of personal information.

Acknowledgments: The author thanks Escuela Politécnica Nacional for the computational resources provided and the technical support of the Faculty of Systems Engineering for conducting the experiments and processing the data.

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

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BETO	Spanish BERT model
CRISP-DM	Cross-Industry Standard Process for Data Mining
F1	F1-score
IA	Artificial Intelligence
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
LLaMA	Large Language Model Meta AI
mBERT	Multilingual BERT
MLM	Masked Language Modeling
NLP	Natural Language Processing
NSP	Next Sentence Prediction
PLM	Pretrained Language Model
ROC	Receiver Operating Characteristic

References

1. Dhanaraj, A. The Evolution of Cyber Threats: From Traditional Attacks to AI-Powered Challenges. *Eur. J. Comput. Sci. Inf. Technol.* **2025**, *13*, 50–61. [CrossRef]
2. Leymann, H. The Content and Development of Mobbing at Work. *Eur. J. Work. Organ. Psychol.* **1996**, *5*, 165–184. [CrossRef]
3. Laboy-Vélez, L.; Ríos-Steiner, A.I.; Flores-Suárez, W. La violencia digital como amenaza a un ambiente laboral seguro. *Forum Empres.* **2021**, *26*, 99–112. [CrossRef]
4. Plaza-del Arco, F.; Molina-González, M.; Ureña López, L.; Martín-Valdivia, M. Integrating Implicit and Explicit Linguistic Phenomena via Multi-task Learning for Offensive Language Detection. *Knowl.-Based Syst.* **2022**, *258*, 109965. [CrossRef]
5. Mabula, M.; Mambeti, S.; Mbelwa, J. Cyberbullying Detection: Exploring Datasets, Technologies, and Approaches on Social Media Platforms. *arXiv* **2024**, arXiv:2407.12154.
6. Báez, P.; Villena, F.; Durán, M. Detecting Cyberbullying in Spanish Texts throughout Deep Learning Techniques. *Int. J. Data Min. Model. Manag.* **2022**, *14*, 234–247 . [CrossRef]
7. Plaza-del Arco, F.; Casavantes, M.; Escalante, H.; Martín-Valdivia, M.; Montejo-Ráez, A.; Montes, M.; Jarquín-Vásquez, H.; Villaseñor-Pineda, L. Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants. *Proces. Leng. Nat.* **2021**, *67*, 183–194.
8. Plaza-del Arco, F.; Molina-González, M.; Ureña López, L.; Martín-Valdivia, M. Comparing Pre-trained Language Models for Spanish Hate Speech Detection. *Expert Syst. Appl.* **2021**, *166*, 114120. [CrossRef]
9. Akosa, J. Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. In Proceedings of the SAS Global Forum 2017, Cary, NC, USA, 2–5 April 2017; Paper 942-2017.
10. Dai, H.; Liu, Z.; Liao, W.; Huang, X.; Cao, Y.; Wu, Z.; Zhao, L.; Xu, S.; Zeng, F.; Liu, W.; et al. AugGPT: Leveraging ChatGPT for Text Data Augmentation. *IEEE Trans. Big Data* **2025**, *11*, 907–918. [CrossRef]
11. Jahan, M.; Oussalah, M.; Beddia, D.; Mim, J.; Arhab, N. A Comprehensive Study on NLP Data Augmentation for Hate Speech Detection: Legacy Methods, BERT, and LLMs. *arXiv* **2024**. [CrossRef]
12. Zambrano, P.; Torres, J.; Anchundia, C.; Illicachi, J. Mobbing: AI-Powered Cyberthreat Behavior Analysis and Modeling. In *Proceedings of the 2024 8th Cyber Security in Networking Conference (CSNet)*; IEEE: Piscataway, NJ, USA, 2024; pp. 278–281. [CrossRef]
13. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.; Kang, H.; Pérez, J. Spanish Pre-Trained BERT Model and Evaluation Data. In Proceedings of the PML4DC at International Conference on Learning Representations 2020, Virtual Conference, 26 April 2020.
14. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
15. Bruns, A. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Inf. Commun. Soc.* **2019**, *22*, 1544–1566. [CrossRef]
16. Plaza, L.; Carrillo-de Albornoz, J.; Arcos, I.; Rosso, P.; Spina, D.; Amigó, E.; Gonzalo, J.; Morante, R. *Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2026; Volume 16089, pp. 266–289.
17. Maqbool, N. *Sexism Identification in Social Networks: Advances in Automated Detection—A Report on the Exist Task at CLEF*; CEUR Workshop Proceedings; CEUR-WS.org: Aachen, Germany, 2024; Volume 3740, pp. 1098–1106.
18. Urios Alacreu, E.; Rosso, P. Identification of Racial and Sexist Stereotypes in Spanish: A Learning with Disagreements Approach. *Proces. Del Leng. Nat.* **2025**, *74*, 15–31.
19. Kharitonova, K.; Pérez-Fernández, D.; Gutiérrez-Hernando, J.; Gutiérrez-Fandiño, A.; Callejas, Z.; Griol, D. EsCorpiusBias: The Contextual Annotation and Transformer-Based Detection of Racism and Sexism in Spanish Dialogue. *Future Internet* **2025**, *17*, 340. [CrossRef]
20. Castillo-López, G.; Riabi, A.; Seddah, D. Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection. In Proceedings of the 10th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Dubrovnik, Croatia, 5 May 2023; pp. 1–13.
21. Mitchell, R. *Web Scraping with Python: Collecting More Data from the Modern Web*, 2nd ed.; O'Reilly Media: Sebastopol, CA, USA, 2018.
22. Yoo, K.; Park, D.; Kang, J.; Lee, S.W.; Park, W. GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event/Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2225–2239.
23. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Proceedings of the Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.

24. Lee, K.; Ippolito, D.; Nystrom, A.; Zhang, C.; Eck, D.; Callison-Burch, C.; Carlini, N. Deduplicating Training Data Makes Language Models Better. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 8424–8445. [CrossRef]
25. Zambrano, P.; Torres, J.; Anchundia, C.; Illicachi, J. Gender-Based Violence as a Cyberthreat: The Impact of Social Media on Digital Security. In *Computer Science and Computational Intelligence (CSCI 2024)*; Communications in Computer and Information Science (CCIS); Springer: Cham, Switzerland, 2025; Volume 2509, pp. 237–256. [CrossRef]
26. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual is Multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4996–5001. [CrossRef]
27. Nozza, D.; Bianchi, F.; Hovy, D. What the [MASK]? Making Sense of Language-Specific BERT Models. *arXiv* **2020**. [CrossRef]
28. Ponnusamy, K.; Vegupatti, M.; Kumaresan, P.; Priyadharshini, R.; Buitelaar, P.; Chakravarthi, B. VEL@IberLEF 2024: Hope Speech Detection in Spanish Social Media Comments using BERT Pre-trained Model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024)*; CEUR Workshop Proceedings; CEUR-WS.org: Aachen, Germany, 2024; Volume 3756.
29. García-Díaz, J.A.; Jiménez-Zafra, S.M.; Valencia-García, R. UMUTeam at HOMO-MEX 2023: Fine-tuning Large Language Models integration for solving hate-speech detection in Mexican Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*; CEUR Workshop Proceedings; CEUR-WS.org: Aachen, Germany, 2023; Volume 3496.
30. Sanchez-Gomez, J.M.; Batista, F.; Vega-Rodríguez, M.A.; Pérez, C.J. A transformer-based deep learning approach for detecting online hate speech in Spanish. *Appl. Soft Comput.* **2026**, *187*, 114259. [CrossRef]
31. Hosseini, H.; Kannan, S.; Zhang, B.; Poovendran, R. Deceiving Google's Perspective API Built for Detecting Toxic Comments. In Proceedings of the 2017 Points of View in Computer Vision Workshop, Venice, Italy, 22–29 October 2017.
32. Grandini, M.; Bagli, E.; Visani, G. Metrics for multi-class classification: An overview. *arXiv* **2020**, arXiv:2008.05756.
33. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; Volume 11.
34. Haixiang, G.; Yijing, L.; Jennifer, S.; Mingyun, G.; Yuanyuan, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
35. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. Predicting the Type and Target of Offensive Posts in Social Media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019.
36. Barbieri, F.; Camacho-Collados, J.; Neves, L.; Espinosa-Anke, L. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020.
37. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
38. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
39. Gholamy, A.; Kreinovich, V.; Olaya, O. *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation*; Technical Report UTEP-CS-18-09; University of Texas at El Paso: El Paso, TX, USA, 2018.
40. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2305016120. [CrossRef] [PubMed]
41. Heaton, J. *Genetic Programming and Evolvable Machines*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 19. [CrossRef]
42. Kaufman, S.; Rosset, S.; Perlich, C. Leakage in Data Mining: Formulation, Detection, and Avoidance. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and 845 Data Mining, San Diego, CA, USA, 21–24 August 2011; Volume 6, pp. 556–563. [CrossRef]

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