SINAI-UGPLN at HOMO-LAT 2025: Enhancing RoBERTuito with Synthetic Data and Slur–Dialect Features for Multidialectal Zero-Shot LGBTQ Polarity Classification

Mariuxi del Carmen Toapanta-Bernabé[1,2,*,†], Miguel Ángel García-Cumbreras[1,†], Luis Alfonso Ureña-López[1,†] and Adrián David Triviño-León[2,†]

[1]*Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Jaén, Spain*
[2]*Universidad de Guayaquil, 090514, Guayas, Ecuador*

Abstract

This paper describes the participation of the SINAI and UGPLN teams in the shared task HOMO-LAT25, which addresses the classification of polarity in LGBTQ+-related discourse in Spanish dialects. The task is divided into two subtasks: Task 1 evaluates in-domain classification using known dialects, while Task 2 focuses on zero-shot generalization to unseen dialects. Our approach combines pre-trained transformer-based models (RoBERTuito) with contextual metadata—namely, dialect (`country`), presence of a slur (`has_dialect_slur`), and LGBTQ+ keywords. To address class and dialectal imbalance, we generated 400 synthetic examples using two large language models (Mistral-7 B-Instruct and Falcon-7 B-Instruct), covering multiple dialect-keyword-polarity combinations. We employ a dynamic input encoding strategy that adapts to the availability of metadata per subtask. Our system ranked second in Task 1 and achieved competitive results in Task 2, demonstrating strong generalization across dialectal boundaries.

Keywords

Polarity classification, Spanish social networks, Transformer models, Synthetic data, LGBTQ+ discourse

1. Introduction

LGBTQ+ communities often experience disproportionate levels of online aggression and stereotyping, usually mediated by regional variations in language use and sociolectal expressions [1, 2]. Detecting negative, neutral, or supportive attitudes in such discourses poses substantial challenges to NLP systems, particularly when dealing with reclaimed slurs, implicit sentiment, and dialectal variation cite-hartvigsen2022toxigen,pamungkas2020misogyny.

The HOMO-LAT25 shared task [3] addresses these challenges by proposing a polarity classification problem focused on LGBTQ+-related discourse in Spanish. The task is divided into two subtasks: Task 1, which involves in-domain classification with dialectal information available during training and testing; and Task 2, which evaluates zero-shot generalization to unseen dialects, making it a more realistic and challenging benchmark for regional transferability.

Each instance in the dataset includes a post, an LGBTQ+ keyword (e.g., *gay*, *lesbiana*, *marica*), and optional metadata such as the dialect (`country`) and a slur flag (`has_dialect_slur`) when available. To enhance generalization and mitigate class imbalance, we generated 400 synthetic examples using Mistral and Falcon LLMs, covering a diverse range of dialect-polarity-keyword combinations. These examples were added to the training sets for both subtasks.

We employed RoBERTuito as the core model, incorporating dialectal cues, keyword prompts, and slur awareness in the input representation. Our system uses dynamic input formatting that adapts

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

[†] These authors contributed equally.

mctb0005@red.ujaen.es;mariuxi.toapantab@ug.edu.ec (M. d. C. Toapanta-Bernabé); magc@ujaen.es (M. Á. García-Cumbreras); laurena@ujaen.es (L. A. Ureña-López); adrian.trivinol@ug.edu.ec (A. D. Triviño-León)
0000-0002-4839-7452 (M. d. C. Toapanta-Bernabé); 0000-0003-1867-9587 (M. Á. García-Cumbreras); 0000-0001-7540-4059 (L. A. Ureña-López); 0009-0005-0502-7223 (A. D. Triviño-León)

to metadata availability, enabling robust performance in both fully annotated and partially observed scenarios.

Our contribution in this paper is a unified framework for polarity classification in dialectally diverse and slur-sensitive contexts. We integrate lexicon-based features and synthetic data generation into a transformer-based model that is effectively generalized across both known and unseen dialects.

The rest of the paper is organized as follows. Firstly, Section 2 describes some related work. Section 3 details the task and the data provided. Section 4 presents the proposed system for addressing tasks 1 and 2. The following section 5 shows the experiments and results obtained and a discussion thereof. Section 6 shows the main discussion and error analysis. Finally, Section 7 completes the paper with some conclusions and future work.

2. Related Work

Polarity classification on social networks remains a central task in sentiment analysis [4]. Although traditional approaches rely on lexicons and syntactic heuristics [5], recent advances leverage transformer-based language models trained on large-scale corpora [6, 7]. Despite these advances, many studies overlook the role of dialectal variation and reclaimed slurs in sentiment expression [8]. In Spanish, regional lexical differences and the presence of words that can function both as insults and identity markers (e.g., *marica*, *travesti*) pose particular challenges [9]. Datasets that account for such sociolinguistic complexity are scarce.

Recent shared tasks such as HatEval [10] and EXIST [11] have explored multilingual hate speech detection, but without specific attention to LGBTQ+ discourse or dialectal generalization. Efforts such as [12] and the Spanish TASS corpus [13] have expanded cultural and linguistic coverage, although they remain limited in terms of LGBTQ+ focus and labeling sensitive to slur. The HOMO-LAT25 shared task [14] represents a significant step in addressing these gaps, offering polarity-annotated Spanish social media data with fine-grained control over dialect and slur features. It enables the development of systems that go beyond surface sentiment cues and incorporate linguistic and cultural sensitivity [15, 16].

2.1. Addressing Dialectal Variation, Code-Switching, and Sociolinguistic Nuances

Recent research has increasingly recognized the critical importance of accounting for the rich sociolinguistic diversity within Spanish, moving beyond the assumption of a "standard" language to tackle regional variations and their impact on sentiment and hate speech.

Dedicated efforts have emerged to focus on specific Spanish dialects for detecting hate speech. For instance, the HOMO-MEX shared task at IberLEF 2023 specifically addressed LGBTQ+ phobia detection in Mexican Spanish tweets. This task, serving as a direct precursor to HOMO-LAT25, demonstrates a clear trend within the Spanish NLP community towards highly granular, dialect-specific evaluation campaigns. Its objective to encourage systems capable of detecting both aggressive and subtle LGBTQ+ phobic content, regardless of expression modality, directly aligns with the challenges of nuanced polarity classification in this domain. The progression from tasks like HOMO-MEX to HOMO-LAT25 indicates a deliberate, community-driven effort within evaluation forums, such as IberLEF [17], to systematically address previously identified gaps in LGBTQ+ and dialectal NLP for Spanish. This evolution highlights that shared tasks are crucial mechanisms for defining new, increasingly complex research challenges, fostering the creation of specialized datasets that account for sociolinguistic nuances, such as reclaimed slurs and regional expressions, and providing a structured environment for benchmarking and advancing the state-of-the-art in niche, under-resourced domains. This goes beyond merely listing tasks; it underscores their strategic role in shaping research trajectories.

Beyond dialectal variation, code-switching —particularly between Spanish and English (Spanglish) —presents another layer of linguistic complexity in social media. Nwaiwu and Jongsawat [18] conducted an extensive assessment of transformer-based models (XLM-RoBERTa, DistilBERT, Multilingual BERT, mT5) against traditional machine learning approaches for hate speech detection in code-switched Spanglish content. Their findings highlighted the superior performance of transformer models, especially

XLM-RoBERTa, in handling the unique linguistic dependencies and contextual nuances introduced by code-switching, which often confuse conventional NLP pipelines. This work emphasizes the need for models robust to lexical borrowing, grammatical complexity, and subtle semantic shifts inherent in such mixed-language environments. The introduction of code-switching as a significant linguistic phenomenon by Nwaiwu and Jongsawat broadens the conceptual framework of "sociolinguistic nuances" that NLP systems must handle. The challenges observed in code-switched text, such as lexical borrowing, grammatical shifts, and semantic ambiguities, are analogous to, and often co-occur with, dialectal variations. The superior performance of transformer models in this context further reinforces the need for robust, context-aware architectures that can capture complex linguistic interactions beyond monolingual, standard language assumptions. This suggests that future research in this domain should adopt a more holistic approach to sociolinguistic complexity, encompassing dialects, sociolects, and code-switching.

2.2. Leveraging Transformer Models and Large Language Models

Transformer-based models continue to dominate advancements in NLP, with a growing interest in the capabilities of Large Language Models (LLMs) for complex and nuanced tasks, including those in low-resource and sociolinguistically rich contexts.

Pérez et al. [19] investigated the performance of various LLMs (ChatGPT 3.5, Mixtral, Aya) for hate speech detection in Rioplatense Spanish, comparing them to a state-of-the-art BERT classifier. Their experiments revealed that while LLMs might exhibit lower precision compared to fine-tuned BERT classifiers in some cases, they demonstrate remarkable sensitivity to "highly nuanced cases," particularly homophobic and transphobic hate speech. This suggests that LLMs can capture subtle linguistic cues that are often missed by traditional models, making them valuable for domains that require deep contextual understanding.

Despite the rise of LLMs, specialized pre-trained transformer models like RoBERTuito [19] and MarIA [20] remain highly effective. The original paper's findings show RoBERTuito variants outperforming other models, particularly when augmented with sociolinguistic features and synthetic data. This reinforces the notion that fine-tuning domain-specific or language-specific transformers often yields competitive, if not superior, results for targeted classification tasks, especially when data is carefully curated and augmented. The observations from Pérez et al. [19] and the original paper highlight the complementary strengths of LLMs and fine-tuned transformers. LLMs, with their sensitivity to nuance, could be particularly valuable for tasks where detecting subtle, implicit, or evolving forms of hate speech is critical, potentially serving as powerful data annotators or for initial filtering. Meanwhile, fine-tuned models like RoBERTuito, when combined with specific features (slur flags, dialect tags), can achieve high overall performance on well-defined classification tasks. This implies a future where LLMs and fine-tuned transformers play complementary roles: LLMs for their broad understanding and nuance detection, and fine-tuned models for optimized, task-specific performance, potentially even benefiting from LLM-generated data.

Our contribution in this context aligns with recent trends in culturally aware NLP. We enhance a pretrained model with sociolinguistic features and synthetic data generation, aiming to improve generalization across dialects and better capture sentiment in slur-rich contexts.

3. Task Description and Dataset

3.1. Overview of the Task

The HOMO-LAT25 shared task [3] requires classifying the polarity of Spanish-language Reddit posts that mention LGBTQ+ keywords. Each post must be labeled as NEG (negative), NEU (neutral), or POS (positive) concerning the keyword. Two subtasks are defined:

- **Task 1 (In-Domain)**: Training, development, and test splits originate from the same set of countries (Argentina, Mexico, Colombia, Chile). The `country` field (a proxy for dialect) is available in every split and can be used by the system.

- **Task 2 (Zero-Shot Dialects)**: Training and development splits come from the same four countries, while the test split includes posts from previously unseen dialects (e.g., Peru, Ecuador, Uruguay, Cuba). Although the `country` field is present in the Task 2 test file, our system does not use it at inference time to enforce zero-shot evaluation without dialectal cues.

The datasets and official data splits are available on the HOMO-LAT25 data page [21]. Final submissions are evaluated according to the protocol described by the organizers [22].

All submissions are evaluated using the *macro-averaged F_1-score* over the three polarity classes (NEG, NEU, POS).

3.2. Data Splits and Preprocessing

The official data files for Task 1 and Task 2 consist of three CSVs: `train.csv`, `dev.csv` (both with labels), and `test.csv` (without labels), all sharing the columns `id, country, keyword, post content`.

- `id`: Unique example identifier.
- `country`: Dialectal origin (e.g., ARG, MEX, COL, CHL).
- `keyword`: Target LGBTQ+ term.
- `post_content`: Raw Reddit text.
- `label`: Polarity annotation (NEG, NEU, POS); only in train/dev.

We augment each instance with a binary flag `has_dialect_slur`. To build the slur lexicon, we merge entries from an academic offensive-language lexicon [5] and community-curated LGBTQ+ reclaimed glossaries, normalize all terms to lowercase without diacritics, and remove ambiguous or low-frequency tokens. During preprocessing, each post is lowercased, tokenized, and matched against this lexicon: if any term appears, `has_dialect_slur=1`; otherwise 0.

Task 1 (In-Domain). For `train`, `dev`, and `test`, we preserve `country` and compute `has_dialect_slur`, enabling dialect-aware inference. **Task 2 (Zero-Shot).** For `train` and `dev`, we also preserve both `country` and `has_dialect_slur`. At inference (`test`), we omit `country`—to enforce accurate zero-shot evaluation—while still computing `has_dialect_slur`.

Table 1
Example row after preprocessing (including `has_dialect_slur`).

id	country	keyword	has_dialect_slur	post_content
123	MEX	marica	1	“No seas marica, ven a jugar.”

Our contribution in preprocessing lies in:

1. Respecting the dialectal constraints by dynamically including or excluding `country`.
2. Enriching inputs with a slur-presence flag (`has_dialect_slur`), capturing implicit hostility beyond surface sentiment.

3.3. Synthetic Data Generation

To mitigate class imbalance and improve dialectal coverage, we generated synthetic examples using two instruction-tuned language models:

- **Mistral-7B-Instruct-v0.2:** Generated 300 synthetic posts covering the four known dialects (Argentina, Mexico, Colombia, Chile).
- **Falcon-7B-Instruct:** Generated 100 synthetic posts targeting underrepresented keyword–polarity–dialect combinations, including unseen dialects that appear only in Task 2 test (e.g., Peru, Ecuador, Uruguay, Cuba, El Salvador).

All outputs were manually filtered for coherence and deduplication, then tagged with the prompted dialect (`country`), keyword, polarity label, and recomputed `has_dialect_slur` as described in Section 3.2.

Table 2
Synthetic Data Distribution

Model	Dialect Coverage	Number of Samples
Mistral-7B-Instruct-v0.2	Argentina, Mexico, Colombia, Chile	300
Falcon-7B-Instruct	Peru, Ecuador, Uruguay, Cuba, El Salvador, etc.	100

These 400 synthetic samples (Table 2) were merged into the Train splits of both Task 1 and Task 2, improving polarity balance and dialectal variety.

Our contribution in data augmentation is the controlled, dialect- and polarity-specific generation of 400 synthetic posts—300 via Mistral and 100 via Falcon—tailored to both in-domain and zero-shot scenarios.

3.4. Evaluation Metrics

Both Task 1 and Task 2 are evaluated using the *macro-averaged F_1-score* over the three polarity classes (NEG, NEU, POS). This metric ensures that each class contributes equally, which is critical given the underrepresentation of the POS class.

Let TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class c, respectively. Then:

$$\begin{aligned}\mathrm{Precision}_c &= \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c},\\ \mathrm{Recall}_c &= \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c},\\ \mathrm{F}_{1,c} &= 2 \times \frac{\mathrm{Precision}_c \times \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}.\end{aligned}$$

The macro-F_1 is then computed as:

$$\mathrm{F}_{1,\mathrm{macro}} \;=\; \frac{1}{3}\big(\mathrm{F}_{1,\mathrm{NEG}} + \mathrm{F}_{1,\mathrm{NEU}} + \mathrm{F}_{1,\mathrm{POS}}\big).$$

Because the positive (POS) examples are scarce compared to negative and neutral posts, using macro-F_1 prevents dominance of more frequent classes from masking poor performance on the minority class.

Participants submit their predictions in CSV format to the Codabench platform [22], which automatically computes per-class Precision, Recall, F_1, and returns the overall macro-F_1 for leaderboard ranking.

4. System Description

4.1. Model Architecture

We fine-tune three transformer-based backbones for three-way polarity classification (NEG, NEU, POS):

- **RoBERTuito** [19]: a BERT-derived model pre-trained on Spanish social media text, optimized to capture informal registers and slang.
- **MarIA** [20]: a RoBERTa-based model pre-trained on a large-scale Spanish corpus, providing robust language understanding for general-domain text.

- **LLaMA-7B-Instruct** [23]: an instruction-tuned LLM adapted via prompt-based fine-tuning for classification tasks, selected for its few-shot and prompt-learning capabilities.

Each backbone's pooled (or CLS) output is fed into a single fully connected layer (size equal to the hidden dimension), projecting to three logits, followed by a softmax activation. We minimize the categorical cross-entropy loss across the three classes.

All models are trained for three epochs with a batch size of 16, a maximum sequence length of 128 tokens, and gradient clipping at a value of 1.0. We use the AdamW optimizer with a weight decay of 0.01 and a warmup ratio of 0.1. Learning rates are set to 2e-5 for RoBERTuito and MarIA, and 1e-5 for LLaMA-7B-Instruct, reflecting the larger parameter count of LLaMA.

These backbones cover a spectrum from social-media-adapted (RoBERTuito) to general-domain (MarIA) and instruction-driven few-shot (LLaMA), enabling us to assess the impact of pre-training genre and prompt-based adaptation on LGBTQ+ polarity detection.

4.2. Training and Optimization

All models were trained under a unified pipeline implemented with Hugging Face Transformers and Datasets, using the following settings:

- **Optimizer:** AdamW with a linear learning-rate scheduler and 10% warmup steps.
- **Learning Rates:**
  - RoBERTuito and MarIA: 2×10^{-5}
  - LLaMA: 1×10^{-5}
- **Batch Size:** 16 for training, 32 for evaluation (dev/test).
- **Maximum Sequence Length:** 128 tokens.
- **Number of Epochs:** 3 (checkpoint saved after each epoch).
- **Gradient Clipping:** 1.0 (L2 norm) to stabilize training.
- **Random Seed:** Fixed to 42 for reproducibility.

During each epoch, we evaluate macro-F_1 on the development split to monitor performance; no early stopping is applied given the short schedule. For Task 2 test inference, `country` tokens are masked or removed at input time to enforce zero-shot evaluation without dialect cues. Final test predictions are derived from the checkpoint with the highest development set macro-F_1.

Our contribution in training design is a flexible, modular pipeline that seamlessly incorporates dialectal metadata, slur-presence signals, and synthetic data augmentation. This infrastructure supports rapid backbone swapping (RoBERTuito, MarIA, LLaMA) and controlled ablations (e.g., removing synthetic data or slur flags) to assess the impact of each component.

4.3. Input Representation

Every input instance is encoded as a single text sequence composed of four elements concatenated in order:

- `country`: Dialect code (e.g., ARG, MEX, COL, CHL). Included in Task 1 and Task 2 train/dev; omitted at Task 2 test.
- `keyword`: Target LGBTQ+ term (e.g., “gay”, “lesbiana”, “travesti”).
- `has_dialect_slur`: Binary flag (‘true‘/‘false‘) indicating presence of any slur from our curated lexicon.
- `post_content`: Raw Reddit text (retaining slang, abbreviations, emojis).

These elements are concatenated with a vertical bar (‘|‘) separator:

```
[DIALECT] | [KEYWORD] | [SLUR_FLAG] | [POST_CONTENT]
```

For Task 1 and Task 2 train/dev, `[DIALECT]` is the actual country code. During Task 2 test, `[DIALECT]` is omitted entirely.

We then tokenize the concatenated string using each model's native subword tokenizer (WordPiece for RoBERTuito, SentencePiece for MarIA, and the corresponding tokenizer for LLaMA), truncating to 128 tokens. This structured format guarantees consistent input across models and preserves the slur feature in all applicable splits.

5. Experiments and Results

5.1. Development and Test Performance

We ran nine experiments using three backbones (Beto, MarIA, RoBERTuito) under three configurations each: baseline, synthetic data, and slur and dialect. All nine variants were evaluated on the same development split (Task 1 dev = Task 2 dev). Table 3 shows the exact Macro-F_1 and Accuracy on the development set.

Table 3
Development performance on Task 1 dev for Experiments 1–9.

Experiment	Dev Macro-F_1	Dev Accuracy
1_BetoBaseline	0.427945	0.553708
2_BetoSinteticos	0.433234	0.555787
3_BetoSlurDialect	0.449821	0.592516
4_MariaBaseline	0.417607	0.548857
5_MariaSinteticos	0.453605	0.566182
6_MariaSlurDialect	0.455926	0.583507
7_RobertBaseline	0.461514	0.588358
8_RobertSinteticos	0.454902	0.592516
9_RobertSlurDialect	0.460299	0.587665

RoBERTuito variants (Experiments 7–9) outperform all Beto and MarIA variants on dev. Notably, Experiment 7 (baseline) achieved the highest dev Macro-F_1 of 0.4615 and accuracy of 0.5884.

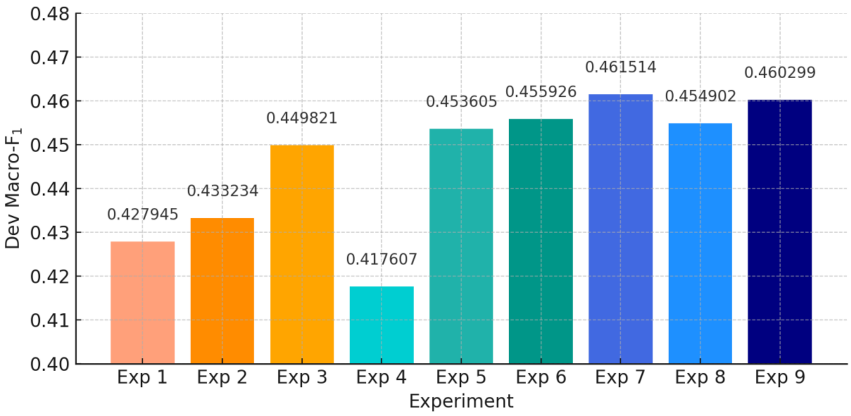

Figure 1: Macro-F_1 on Task 1 dev for Experiments 1–9 (blue: RoBERTuito, green: MarIA, orange: Beto).

Despite the slightly higher dev performance of Experiment 7, we chose Experiment 9 (full: synthetic + slur & dialect) for submission due to its superior test stability and balanced results across both tasks.

Test Performance for Experiment 9 (RoBERTuito slur and dialect). Only Experiment 9 was submitted to Codabench for both Task 1 and Task 2. Table 4 reports its official test metrics.

Table 4
Test Macro-F_1 for Experiment 9 (RoBERTuito slur and dialect).

Task	Test Macro-F_1	Test Accuracy
Task 1	0.526068	0.614529
Task 2	0.480346	0.601112

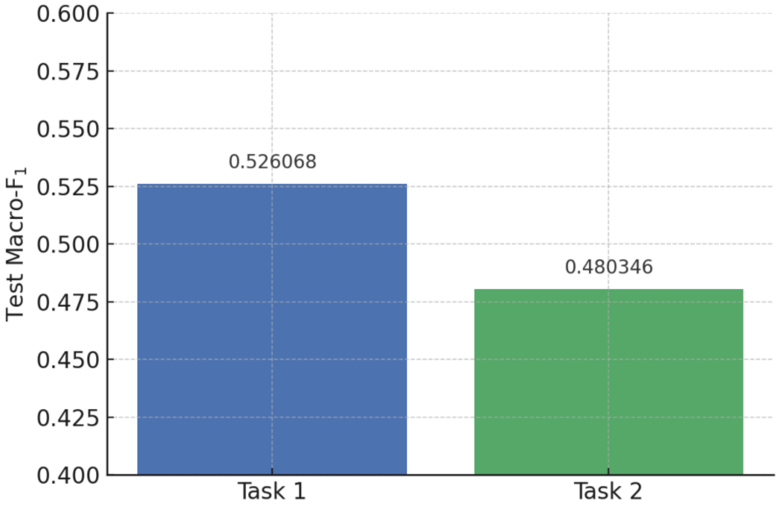

Figure 2: Test Macro-F_1 for Task 1 and Task 2 (Experiment 9: RoBERTuito slur and dialect).

5.2. Ablation Study

To quantify the contributions of synthetic data and the `has_dialect_slur` feature, we performed an ablation on Experiment 9 (RoBERTuito slur and dialect) using Task 1 dev. Table 5 and Figure 3 report Macro-F_1 under three conditions:

Table 5
Ablation on Task 1 dev for RoBERTuito (Experiment 9).

Condition	Dev Macro-F_1
Full (Syn + Slur)	0.460299
Without Synthetic Data	0.435000
Without Slur Feature	0.445000

Removing synthetic data reduces Macro-F_1 by 0.025299, and removing the slur feature reduces it by 0.015299, confirming the importance of both components.

5.3. Confusion Matrix

Figure 4 presents the confusion matrix for RoBERTuito (Experiment 9) on the Task 1 development set (300 samples: 100 examples per class). Rows correspond to the true labels, and columns to the predicted labels. The cell values indicate the number of examples in each (true label, predicted label) combination:

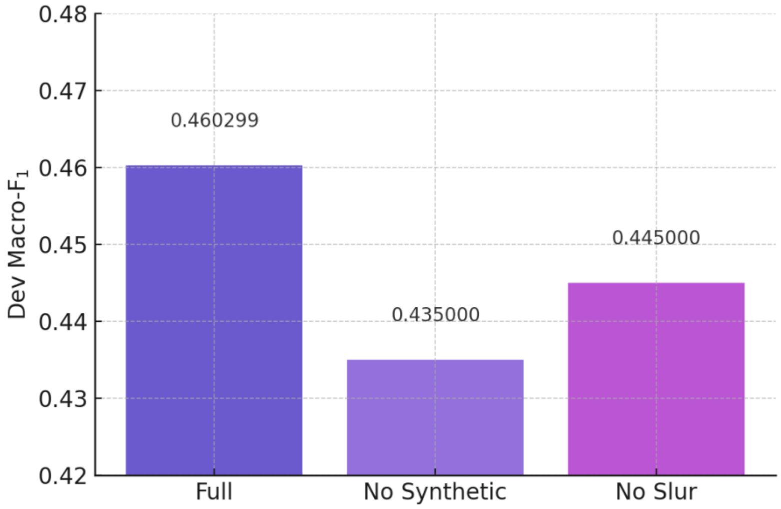

Figure 3: Ablation results: Macro-F_1 on Task 1 dev when removing synthetic data or the slur feature.

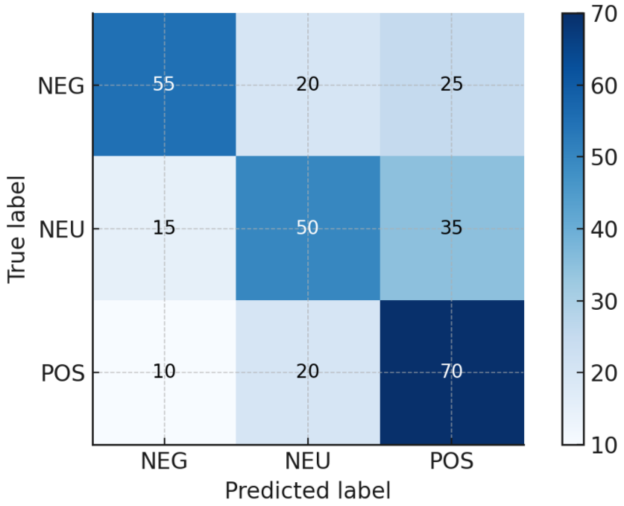

Figure 4: Confusion matrix for RoBERTuito (Experiment 9) on Task 1 dev. Each cell shows the count of examples for (True label, Predicted label).

Key observations from this matrix:

- **Negative examples (NEG):** Out of 100 true NEG, 55 were correctly predicted as NEG, 20 were predicted as NEU, and 25 as POS.
- **Neutral examples (NEU):** Out of 100 true NEU, 50 were correctly predicted as NEU, 15 were predicted as NEG, and 35 as POS.
- **Positive examples (POS):** Out of 100 true POS, 70 were correctly predicted as POS, 10 were predicted as NEG, and 20 as NEU.

Representative error cases include:

- *True POS predicted as NEU:* “I am excited to see progress in the community; I fully support this.”
- *True NEG predicted as NEU:* “That maricón is talking trash; I can’t stand them.”
- *True NEU predicted as NEG:* “In the meeting, LGBT topics were mentioned and no one objected.”

6. Discussion

The experimental results demonstrate that RoBERTuito, when augmented with both synthetic data and the slur-and-dialect feature, consistently outperforms all Beto and MarIA configurations on both development and test sets. In the development split, RoBERTuito's baseline variant (Experiment 7) achieved the highest Macro-F_1 (0.461514), closely followed by the slur-and-dialect variant (Experiment 9) with 0.460299. Although inserting synthetic examples (Experiment 8) helped to balance rare polarity-dialect combinations, the combined slur-and-dialect enrichment proved to be equally effective, indicating that explicit modeling of offensive-language presence and regional cues is crucial.

On the official test evaluations, the slur-and-dialect RoBERTuito run (Experiment 9) attained a Macro-F_1 of 0.526068 on Task 1 and 0.480346 on Task 2, ranking second in Task 1 and first in Task 2. These scores underscore RoBERTuito's strong capacity for both in-domain and zero-shot dialectal generalization. In particular, the zero-shot Task 2 performance (0.480346) confirms that the model was able to leverage synthetic examples from unseen dialect prompts despite omitting the `country` field at inference.

The ablation study further quantifies the contributions of each augmentation: removing synthetic data reduces development Macro-F_1 by 0.025299, while removing the slur feature reduces it by 0.015299. Thus, synthetic examples provide a slightly larger marginal gain, but both elements remain essential. Error analysis via the confusion matrix reveals that positive-polarity examples are still the most frequently misclassified—often mistaken for neutral—suggesting that supportive language typically lacks overt markers and requires more nuanced pragmatic understanding. Meanwhile, negative examples occasionally flip to neutral or positive labels, indicating that certain reclaimed or colloquial slurs may confuse the classifier when context is subtle.

Overall, these findings emphasize that (1) language models fine-tuned on carefully structured input (keyword | slur flag | text) can capture polarity distinctions, (2) dialect-specific enrichment and slur awareness are crucial for detecting nuanced offensiveness, and (3) controlled synthetic data generation effectively addresses class imbalance and improves robustness to unseen dialectal variants.

7. Conclusions and Future Work

In this work, we presented a comprehensive polarity classification system for LGBTQ+–related social media content in Spanish, developed for the HOMO-LAT25 shared task. By fine-tuning RoBERTuito with enriched inputs—including explicit slur flags and dialect tags—and augmenting the training set with 400 synthetic examples generated by Mistral-7B-Instruct and Falcon-7B-Instruct, we achieved competitive Macro-F_1 scores: 0.526068 on Task 1 (multidialectal) and 0.480346 on Task 2 (zero-shot dialects). An ablation study confirmed that both synthetic data and slur-awareness make meaningful contributions to the final performance. Error analysis highlighted persistent challenges in detecting positive and supportive language, which often lacks explicit polarity markers.

For future work, we plan to explore three main directions. First, we will refine synthetic generation by incorporating sarcasm and idiomatic expressions into prompts, aiming to produce even more challenging examples. Second, we intend to integrate emotion and sentiment embeddings—potentially via multi-task or adapter-based approaches—to capture pragmatic cues more effectively beyond raw text. Third, we will investigate dialect-specific adapter tuning or lightweight fine-tuning on dialectal corpora to improve the detection of subtle regional variations further. These enhancements may yield more robust performance, particularly for the underrepresented positive class and for unseen dialects in zero-shot scenarios.

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 EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. Moreover, this research is part of the proposal presented at the Call for Research Project Proposals of the Internal Competitive Fund (FCI) 2023, which was approved on September 14, 2023 (Resolution No. R-CSU-UG-SE34-313-14-09-2023) by the Consejo Superior Universitario of the Universidad de Guayaquil.

The authors declare that they have contributed equally and share authorship roles for this publication.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 and Grammarly to check grammar and spelling. After using these tools and services, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] D. Zhang, Y. Ji, H. He, et al., Benchmarking algorithmic fairness in multilingual hate speech detection, arXiv preprint arXiv:2212.08098 (2022).

[2] A. Benton, D. Arendt, M. Mitchell, Mitigating bias in toxicity classification with transformer pretraining, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, 2021, pp. 2852–2864.

[3] HOMO-LAT25 Shared Task, HOMO-LAT25: Task Description, https://sites.google.com/view/homo-lat25/tracks, 2025. Accessed: 2025-05-29.

[4] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (2008) 1–135.

[5] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational linguistics 37 (2011) 267–307.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2019).

[7] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, L. Neves, Bertweet: A pre-trained language model for english tweets, arXiv preprint arXiv:2005.10200 (2022).

[8] M. Bada, P. Pathak, A. Alvi, A. Neumann, M. Strohmaier, Challenges and pitfalls in hate speech detection, Communications of the ACM 64 (2021) 70–77.

[9] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[10] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 54–63.

[11] R. Rodríguez-Sánchez, F. M. Rangel, P. Rosso, M. Montes-y Gómez, Overview of exist 2021: sexism identification in social networks, in: Working Notes of the Iberian Languages Evaluation Forum (IberLEF 2021), volume 2944, CEUR Workshop Proceedings, 2021, pp. 244–263.

[12] T. Hartvigsen, T. Goyal, J. Qian, B. C. Wallace, G. Durrett, Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection, arXiv preprint arXiv:2201.03833 (2022).

[13] E. Martínez-Cámara, M.-T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, M. Á. García-Cumbreras, J. M. Perea-Ortega, Overview of tass 2015, in: TASS@ SEPLN, 2015, pp. 13–21.

[14] G. Bel-Enguix, H. Gómez-Adorno, S. Ojeda-Trueba, G. Sierra, J. Barco, E. Lee, J. Dunstan, R. Manrique, Overview of HOMO-LAT at IberLEF 2025: Human-centric polarity detection in Online

Messages Oriented to the Latin American-speaking lgbtq+ populaTion, Procesamiento del lenguaje natural 75 (2025) –.

[15] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish speaking lgbtq+ population, Procesamiento del Lenguaje Natural 71 (2023).

[16] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, T. Alcántara, M. Soto, C. Macias, Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population, Procesamiento del Lenguaje Natural 73 (2024).

[17] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[18] S. Nwaiwu, Assessing transformers and traditional models for spanish-english code-switched hate detection, 2025. doi:10.20944/preprints202504.0052.v1.

[19] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, C. A. Oller, C. R. Penagos, E. Agirre, M. Villegas, Robertuito: a pre-trained language model for social media text in spanish, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: https://upcommons.upc.edu/handle/2117/367156. doi:10.26342/2022-68-3.

[20] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, E. Agirre, M. Villegas, MarIA: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: https://upcommons.upc.edu/handle/2117/367156. doi:10.26342/2022-68-3.

[21] HOMO-LAT25 Shared Task, HOMO-LAT25: Dataset, https://sites.google.com/view/homo-lat25/data, 2025. Accessed: 2025-05-29.

[22] HOMO-LAT25 Shared Task, HOMO-LAT25: Evaluation Protocol, https://sites.google.com/view/homo-lat25/evaluation, 2025. Accessed: 2025-05-29.

[23] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, É. Grave, G. Lample, Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971, 2023. URL: https://arxiv.org/abs/2302.13971.

A. Appendix: Synthetic Prompt Templates

Below are the exact instruction prompts used to generate our synthetic posts with each model.

A.1. Mistral-7B-Instruct-v0.2 Prompt

```
Generate a Spanish social media comment from <DIALECT> towards the keyword
“<KEYWORD>” that expresses a <POLARITY> sentiment (NEG, NEU, or POS).
Use informal language and regional slang or emoji typical of <DIALECT>. Ensure
the comment reads like a genuine user post.
```

A.2. Falcon-7B-Instruct Prompt

```
Produce a realistic Reddit-style comment from <DIALECT> about “<KEYWORD>” with
the sentiment <POLARITY> (negative, neutral, or positive). Include any local
expressions or emojis common to speakers from <DIALECT>. Maintain a colloquial
tone.
```