INFOTEC-NLP at MultiPRIDE: Region-Aware Transformer Models for Spanish Reclaimed Language Classification

Jorge Gleaves[1,*,†], Guillermo Ruiz[1,†]

[1] *INFOTEC Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, 112 Circuito Tecnopolo Sur, Parque Industrial Tecnopolo 2, Aguascalientes, 20326, México.*

Abstract

This report describes the system submitted for the MultiPRIDE Evalita 2026 Shared Task, specifically focusing on Task A, Subtask A2: Spanish Reclaimed Language Classification. [1] The task challenges participants to classify whether a term related to the LGBTQ+ context within a Spanish sentence is used with a reclaimed intent or not. This is framed as a binary classification problem. We fine-tuned two language models trained in Spanish, one limited to the Mexican Spanish and the other to Latin America. We achieved F1-scores of 0.7034 and 0.7569 on the test set, respectively. Motivation: The increasing presence of LGBTQ+ communities online has led to the appropriation of previously derogatory terms as symbols of identity and empowerment. Understanding and accurately classifying such ”reclaimed language” is crucial for nuanced sentiment analysis, content moderation, and supporting inclusive online environments. It helps distinguish between hate speech and self-affirming expression, contributing to a more respectful digital discourse. Warning: This paper contains examples of explicitly offensive content.

Keywords

Reclaimed language, Spanish language processing, Binary text classification, Hate speech detection, LGBTQ+

1. Introduction

The rapid growth of user-generated content on social media platforms has intensified the need for robust Natural Language Processing (NLP) methods capable of identifying nuanced forms of meaning [2], intention, and social context. In particular, the analysis of language related to the LGBTQ+ community poses unique challenges, as certain terms historically associated with hate speech or discrimination may be deliberately reappropriated and used with self-affirming or empowering intent. Accurately distinguishing between harmful and reclaimed uses of such language is essential for applications including sentiment analysis, hate speech detection, and content moderation, where superficial lexical cues are often insufficient.

This paper presents the system developed by the team INFOTEC-NLP for participation in the MultiPRIDE Evalita 2026 Shared Task, focusing on Task A, Subtask A2, which addresses reclaimed language classification in Spanish. The task is formulated as a binary classification problem: given a Spanish sentence containing a term related to the LGBTQ+ context, the system must determine whether the term is used with a reclaiming intent or a non-reclaiming (potentially offensive or neutral) intent. The challenge lies in capturing pragmatic, cultural, and contextual signals that influence interpretation, particularly in a language as regionally diverse as Spanish.

1.1. Task description

The task aims to analyze the use of reclaimed language within the LGBTQ+ community. The goal is to determine whether certain terms —historically used as offensive or denigratory— are employed with a reclaiming or self-affirming intent. Participants are encouraged to consider:

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT

[*] Corresponding author.

[†] These authors contributed equally.

jggleaves@gmail.com (J. Gleaves); luis.ruiz@infotec.mx (G. Ruiz)

0000-0001-7422-7011 (G. Ruiz)

- Textual content of the message, such as arguments, slurs, derogatory terms, self-labeling, and figures of speech.
- Contextual information that can be inferred (when available), such as whether the author belongs to the LGBTQ+ community or their political orientation.

The problem is framed as a binary classification task, where systems must decide whether a term related to the LGBTQ+ context is used with reclamatory intent or non-reclamatory intent.

Task A - Textual Content focuses on systems that operate solely on the textual content of the message, without incorporating explicit user profile information. Two methodological approaches are permitted. In the *constrained approach*, participants may use only the training data provided for the task; external resources such as lexicons are allowed, but no additional annotated datasets (e.g., tweets, sentences, or corpora labeled for reclaimed language) may be incorporated. In contrast, the *unconstrained approach* permits the use of supplementary training material, including additional annotated datasets relevant to reclaimed language or other auxiliary resources. Participants choosing to work in the unconstrained setting must clearly indicate this choice when submitting their runs and provide a detailed explanation of the additional data and methods used in their accompanying technical report. **Our team selected the constrained approach.**

Subtask A2 - Spanish (Main Focus)

Participants work exclusively with Spanish texts. The objective is to identify whether LGBTQ+ related terms in a Spanish sentence are used with a reclaiming intent. The main challenge lies in capturing the pragmatic and cultural nuances of Spanish, where the same term may be offensive or reclaiming depending on context, tone, and identity-based usage.

In summary, Subtask A2 evaluates the ability of a system to understand context, intention, and semantic reclamation in Spanish, a task in which surface-level cues are often insufficient and deeper linguistic understanding is required.

To address this task, we explore the fine-tuning of two Transformer-based language models pretrained on large-scale Spanish Twitter corpora with regional awareness [3]: MEX_Large [1], specialized in Mexican Spanish, and BilmaLAT [2], covering multiple Latin American varieties. Both models are adapted to the task using the training data provided by the organizers, following a constrained approach, and were evaluated on the official validation and test sets.

Roadmap

The remainder of this paper is structured as follows. Section 3 describes the proposed system, including data preprocessing, tokenization strategies, model architecture, and hyperparameter optimization. Section 4 reports the experimental results obtained on the validation and test sets. Section 5 discusses the findings, highlights the main challenges observed—particularly the class imbalance and performance, performance—and outlines the directions for future work.

2. Related work

Hate speech detection is a problem that has been introduced in previous challenges, like PAN 2021 [4] which introduced three shared tasks—cross-domain authorship verification, profiling hate-speech spreaders on Twitter, and style-change detection in multi-author documents—designed to push forward reproducible research in text forensics and stylometry through new benchmark datasets and standardized evaluation. In recent years, challenges such as HOMO-MEX (2023) [5] and HOMO-LAT (2025) [6] focused on detecting LGBTQ+-related hate speech and sentiment in Spanish on Twitter and Reddit,

[1] https://huggingface.co/guillermoruiz/MEX_Large
[2] https://huggingface.co/guillermoruiz/BilmaLAT

using annotated corpora and Transformer-based models that show moderate success but struggle with linguistic ambiguity and regional variation. Together, they highlight the need for more robust NLP systems based on dialect - to better classify LGBTQ+-phobic content and polarity in diverse Spanish-speaking communities.

The Transformer architecture [7] has proven to be very effective. One of the main differences with previous methods is the use of the Attention mechanism to understand the context of the words. The BERT and RoBERTa [8] models are good examples of successful models based on Transformers. The main drawback of these pretrained models is that they are focused on the English language and their performance is poor for a different language, like Spanish. This is the motivation for training models from scratch with regional information. The MEX_Large and BilmaLAT models are based on the RoBERTa architecture and were trained on a large collection of georeferenced tweets. The former was trained on messages from Mexico and the later from Latin America countries.

MEX_Large and BilmaLAT models include the option to introduce special regional tokens to guide the prediction, that is, each model has special tokens embedded; for MEX_Large, the token represents a state from Mexico's country, for BilmaLAT, the tokens represent a country from Latin America, and a time window consisting of a year and a month. See Table 1 for some tweet examples. For this challenge, the publication of the date was not necessary; therefore, we decided to fix it at _2023 _01.

For this binary classification task, the models were initialized by adding a classification head appropriate for our two target classes.

Table 1
Examples of tweets and the special tokens for each model. These are examples after the pre-processing.

Token + Tweet	Model
Coahuila _GEO Cómo estás amiga, nos conocemos? Soy soltero busco soltera. #PiedrasNegras #nava #allende #zaragoza	MEX_Large
Tamaulipas _GEO Ando de buenas que ya les devolví sus unfollows y métanselos por el culo ☺.	MEX_Large
BCS _GEO Ésa canción que cantas en silencio y la otra persona tmb. Bn raro.	MEX_Large
_do _2017 _09 Y que es lo que uno va hacer con _usr ? E ponen hoy mi Plan de dato y con 5 minutos de eso se acabó _usr _usr	BilmaLAT
_ar _2020 _03 soy demasiado buena para todo el mundo y se viven cagando en mí, que hago? me vuelvo una forra de mierda así me valoran?	BilmaLAT
_mx _2019 _12 Felicidades para la nueva pareja, y para usted también. Gracias por el pedazo de pastel. _url	BilmaLAT

3. Our system solution

Our approach was to fine-tune an early bird version of our BilmaLAT [3] and MEX_Large [4] models. The models were designed such that the where and when of a given text is captured, i.e. it learns about regions and periods of time; the model can be accessed and tested with the Huggingface framework

The core of our methodology involves fine-tuning a pre-trained multilingual Transformer model using the provided Spanish dataset. We did not use ensambles because we wanted to have a benchmark of our models. It is important to point out that the models are small (130 million parameters) and can be used or fine-tunned using personal computers or free cloud computing services like Colab.

3.1. Data Preprocessing and Cleaning

Before feeding the text into the model, a cleaning function clean_tweet, was applied to the text column of both the training and test datasets. This function performs the following steps:

[3] https://huggingface.co/guillermoruiz/BilmaLAT

[4] https://huggingface.co/guillermoruiz/MEX_Large

1. HTML Tag Removal: Uses BeautifulSoup to remove any HTML/XML tags.
2. User Mention Removal: Removes Twitter-style user mentions (e.g. @USER).
3. URL Removal: Removes URLs from the text.
4. Non-alphabetic Character Removal: Retains only alphabetic characters, periods, exclamation marks, and apostrophes.
5. Whitespace Normalization: Replaces multiple spaces with a single space.

3.2. Tokenization

After a complete cleaning of the data, we created a balanced dataset. The results of the models were average. On the other hand, using the data without any clean process, the results presented an improvement, therefore the tokenization is applied to the dataset without any clean process.

Tokenizer usage. Each model used its corresponding pretrained tokenizer (the RoBERTa-style tokenizers provided with BilmaLAT and MEX_Large). Text was tokenized to the model's maximum sequence length used during fine-tuning.

3.2.1. Region token selection.

- BilmaLAT: we prepended the regional token to each example. This token denotes Spain and was chosen because the dataset was assumed to contain a majority of tweets from Spain; adding provides an explicit regional signal to the multiregional model.
- MEX_Large: we prepended the regional token Mexico_City. This token was selected because Mexico City is treated as a neutral, broadly representative regional variant for Mexican Spanish idioms and thus reduces region-specific idiomatic bias.

Year/month (time) token selection.

- BilmaLAT time-token constraint: BilmaLAT restricts allowed time tokens to the range 2015–2023. To remain within this restriction and to provide a consistent temporal context, we fixed the time tokens to January 2023 (__2023 __01) for all examples. This choice supplies a stable "when" signal without relying on per-example metadata that was not required by the task.
- MEX_Large time tokens: no task-specific time constraint was required for MEX_Large in our setup; we used the regional token only Mexico__City and omitted varying time tokens to avoid introducing noisy temporal variation.

For each input sentence we prepended the chosen region token and the fixed time tokens (where applicable) before tokenization so the model could attend to region/time context from the first token positions.

3.3. Hyperparameter Search

To optimize the performance of the model, an automated hyperparameter search was conducted using Optuna [9]. The Trainer was re-initialized with a model__init function to ensure a fresh model instance for each trial. The hp__space function defined the search range for:

- learning__rate: Log-uniform distribution between 1×10^{-5} and 5×10^{-5}.
- num__train__epochs: Integer values between 4 and 12.
- per__device__train__batch__size: Categorical choices of $[16, 32, 64]$.

Trials and evaluation; We performed 10 Optuna trials per model. Each trial:

1. instantiated a fresh model via model__init,
2. trained with the sampled hyperparameters and
3. evaluated in the validation split; Optuna selected the hyperparameter set with the highest validation F1-macro.

The notebook shows the trainer.hyperparameter__search(..., n__trials=10) call and subsequent extraction of best__run.hyperparameters.

Table 2
Confusion matrix on the development set for BilmaLAT.

	True	False
Label 1, Reclaimed	6	2
Label 0, Non-reclaimed	31	3

Table 3
Classification Report for the validation set using BilmaLAT

	precision	recall	f1-score	support
0	0.9118	0.9394	0.9254	33
1	0.7500	0.6667	0.7059	9
accuracy			0.8810	42
macro avg	0.8309	0.8030	0.8156	42
weighted avg	0.8771	0.8810	0.8783	42

Table 4
Confusion matrix on the development set for MEX_Large

	True	False
Label 1, Reclaimed	6	0
Label 0, Non-reclaimed	33	3

3.4. Final Model Training

After identifying the best hyperparameters, the models were re-initialized from scratch and trained using the optimal parameters found during the Optuna search. The best hyperparameters found for BilmaLAT were: learning rate of 2.6187×10^{-5}, 5 epochs and a batch size of 32. We repeated the same process for MEX_Large and the hyperparameters used were: 1.9998×10^{-5}, 8 epochs and a batch size of 32.

The final training used the provided training set (785 samples) and evaluation was performed on the official validation set (42 samples). Models were trained with standard cross-entropy loss for binary classification; class weighting or specialized loss functions were not applied in the reported runs (but are recommended to mitigate class imbalance).

4. Results

4.1. Validation Set Performance (after Hyperparameter Tuning)

After Optuna tuning we re-trained each model with the selected hyperparameters and evaluated on the official validation split (42 samples). Validation metrics reported below are the model scores obtained after the final training run described in Section 3.4 For the BilmaLAT, we got the results shown on Tables 2 and 3. The model achieved an accuracy of 88.10% and an F1-macro score of 81.56% in the validation set.

For the case of the MEX_Large model, the results are shown in Tables 4 and 5. As can seen, the model achieved an accuracy of 92.86% and an F1-macro score of 87.83% in the validation set.

4.2. Test Set Predictions and Official Results

The final predictions on the held-out test set were produced using the re-trained models and submitted to the shared task evaluation. The official test metrics reported below are those returned by the MultiPRIDE evaluation server for each model. For BilmaLAT results are shown in Table 6 where we can see a score

Table 5
Classification Report for validation set using MEX_Large

	precision	recall	f1-score	support
0	0.9167	1.0000	0.9565	33
1	1.000	0.6667	0.8000	9
accuracy			0.9286	42
macro avg	0.9583	0.8333	0.8783	42
weighted avg	0.9345	0.9286	0.9230	42

Table 6
BilmaLAT Official results

	precision	recall	f1-score
0	0.9372	0.9014	0.9189
1	0.5420	0.6590	0.5948
macro avg	0.7396	0.78024	0.7569

Table 7
MEX_Large Official results

	precision	recall	f1-score
0	0.8948	0.9758	0.9335
1	0.7209	0.3522	0.4732
macro avg	0.8078	0.6640	0.7034

of 0.7569 of macro F1. The results for MEX_Large are shown in Table 7. It scored 0.7034 of macro F1. We can see a clear advantage of the BilmaLAT model over the MEX_Large.

5. Discussion

The performance in the validation set, with an F1-macro score of 75.69% and 70.34%, suggests that the fine-tuned BilmaLAT and MEX__Large models are reasonably effective in identifying reclaimed language in Spanish tweets, however, we can see a struggle for identification of Label 1 corresponding to reclaimed language; some examples are shown in the table 8.

Class Imbalance Impact: The class imbalance in the training data (85% Label 0, 15% Label 1) is a critical factor. Although the model shows high precision and recall for the majority class Label 0, the performance for the minority class Label 1, reclaimed language, has lower precision (0.5420) and recall (0.6590), with an F1-score of 0.5948 for BilmaLAT. This is a common challenge in imbalanced datasets, where models tend to be biased towards the majority class. However, MEX_Large seems to have more problems in identifying claims, showing a slight decrease in general.

The contrast between per-class precision and recall shows two complementary failure modes:

1. Low precision on Label 1 (BilmaLAT test) means that many predicted reclaimed examples are actually non-reclaimed;
2. Low recall on Label 1 (MEX_Large test) means that many true reclaimed examples are missed.

Both reduce macro F1 and are symptomatic of the skewed training distribution and the pragmatic difficulty of inferring reclamation from short social messages.

Hyperparameter Tuning Effectiveness: The Optuna-based hyperparameter search proved beneficial in identifying an optimal combination of learning rate, number of epochs and batch size, which likely contributed to the final validation performance.

Table 8
Examples where the model BilmaLAT selected the wrong label

id	text	Correct label
es_989	Hace dos días para vosotros la ley trans era un “retroceso en derechos” y los que estaban en la calle recibiendo palos ayer enfrentándose a la poli eran trans, maricones y bolleras; tened un poquito de vergüenza y dejad de usarnos para los argumentos que os vienen bien URL	1
es_1343	Este es el mes trending!!!! Pero pa jotear y mariconear esta todo el año y toda la vida!! #queer #pride #loveislove #LGTBI URL	1
es_1394	No importa la edad que tengas, los brazos de una madre siempre son el mejor refugio #FelizDiadeLaMadre #QueerAsFolk #FelizViernes #LoveIsLove #LGBT #LGTBI #Gay #GayBlog #GayKiss #GayLove #RelatoLGBT #LGTBIQ #AmorLGBT #GayPride #AmbienteGay #Orgullo2021 #Gays #Pride2021 URL	0
es_469	Por todos los que han sido humillados desde la infancia. Agredidos o asesinados, al grito de maricón. A los que no quieren que seamos familia. A los que reivindican ”terapias de conversión” A los que no quieren que se exhiba la bandera LGTBI. Visibilicemos más el #OrgulloLGTBI URL	0

Choice of Base Model: BilmaLAT and MEX_Large, being RoBERTa-based models pre-trained on a large Spanish corpus from Twitter (now X), aligns well with the task of identifying nuanced expressions in social media contexts. However, on this occasion, the task was to detect reclaimed language that is used with pride, creating an appropriation of terms for the LGBTQ+ population; this complicated the specific task and resulted in poor results in Label 1; on the other hand, it is shown that the use of special tokens for the regional selection of the Spanish language was a good choice, where BilmaLAT, adjusted for Spanish from Spain, performed better than MEX_Large with adjustments for Spanish from Mexico.

Conclusions and future work

In this work, we fine-tuned two region-aware RoBERTa models (BilmaLAT and MEX_Large) to detect reclaimed language in Spanish social posts. We achieved F1-scores of 75.69% and 70.34% in the test set, respectively. The experiments show consistent strengths on **Label 0** (non-reclaimed) and persistent weaknesses on **Label 1** (reclaimed), a pattern largely explained by the training distribution (85% Label 0, 15% Label 1). Overall, region/time tokens and Optuna tuning improved generalization, but the models still struggle to reliably identify reclaimed uses when pragmatic cues are subtle or removed by aggressive cleaning.

Future work

Our results are a clear sign of the need to have dedicated language models that are aware of regional information. This is an opportunity area for future research on NLP. We identified some improvements that could be applied in order to enhanced performance:

- **Addressing Imbalance**: Future work could explore techniques like oversampling the minority class (e.g., SMOTE), under sampling the majority class, using class weights during training, or employing different loss functions (e.g., focal loss) to mitigate the impact of class imbalance and potentially improve performance on the reclaimed language class.
- **Contextual Features**: The current approach is based primarily on text content. Incorporating contextual information from user profiles (as mentioned in the task description) could provide valuable signals to discern reclaimed intent.

- **Ensemble Methods**: Combining predictions from multiple models or different model architectures could lead to more robust performance.
- **Error Analysis**: A detailed error analysis on misclassified validation samples could re-veal patterns and guide further model improvements or feature engineering.

Declaration on Generative AI

During the preparation of this work, we applied the Writefull's model for grammar and spelling checks. After using these services, we reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

[1] C. Ferrando, L. Draetta, M. Madeddu, M. Sosto, V. Patti, P. Rosso, C. Bosco, J. Mata, E. Gualda, Multipride at evalita 2026: Overview of the multilingual automatic detection of slur reclamation in the lgbtq+ context task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[2] D. Rao, B. McMahan, Natural Language Processing with PyTorch Intelligent Language Applications Using Deep Learning, Oreilly press, 2019.

[3] G. Ruiz, R. Campos, T. Ramirez-delreal, D. Moctezuma, M. Graff, E. S. Tellez, Infotec-nlp at homo-lat 2025: Testing a novel multi-region spanish model to monitor opinion in latin american lgbtqi+ social media (2025).

[4] J. Bevendorff, B. Chulvi, G. L. D. L. P. n. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection: Extended abstract, in: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2021, p. 567–573. URL: https://doi.org/10.1007/978-3-030-72240-1_66. doi:10.1007/978-3-030-72240-1_66.

[5] G. B.-E. y Helena Gómez-Adorno y Gerardo Sierra y Juan Vásquez y Scott Thomas Andersen y Sergio 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) 361–370. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6566.

[6] G. B.-E. y Helena Gómez-Adorno y Sergio Ojeda-Trueba y Gerardo Sierra y Jessica Barco y Edgar Lee-Romero y Jocelyn Dunstan y Ruben 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) 413–424. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6764.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[8] S. Jimenez, G. Dueñas, A. Gelbukh, C. A. Rodriguez-Diaz, S. Mancera, Automatic detection of regional words for pan-hispanic spanish on twitter, in: G. R. Simari, E. Fermé, F. Gutiérrez Segura, J. A. Rodríguez Melquiades (Eds.), Advances in Artificial Intelligence – IBERAMIA 2018, Springer International Publishing, Cham, 2018, pp. 404–416.

[9] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.