LlaNa at MultiPRIDE: A Fine-Tuning Approach with Cross-Lingual Augmentation

Alessio Mercurio[1,†], Giorgio Talluto[1,†], Irene Siragusa[1,*] and Roberto Pirrone[1]

[1]Department of Engineering, University of Palermo, Viale delle Scienze, Edificio 6, Palermo, 90128, Sicily, Italy

Abstract

In this report we presents the proposed approach of LlaNa team for the MultiPRIDE challenge at EVALITA 2026 campaign. We propose a robust fine-tuning pipeline based on XLM-RoBERTa for both Italian tasks A (textual) and B (textual + user bio), enhanced by a targeted cross-lingual data augmentation strategy. By translating reclamation instances from the auxiliary English and Spanish datasets into Italian, we successfully enriched the minority class, enabling the model to learn reclamation patterns more effectively. The experimental results demonstrate the effectiveness of the proposed data-centric strategy. The developed system won in subtask B1 reaching a macro F1 score equal to 0.9021, and placed at the 5th place in subtask A1 with a score equal to 0.8835.

Keywords

LGBTQ+ Slurs Reclamation, Cross-Lingual Augmentation, Language Models

1. Warning

This paper contains examples of explicitly offensive content.

2. Introduction

The proliferation of hate speech on social media platforms has driven the Natural Language Processing (NLP) community to develop increasingly sophisticated automatic detection systems. While significant progress has been made in identifying toxic content, current state-of-the-art models often struggle with the nuances of language use within marginalized communities [1]. A critical limitation of many existing systems is their over-reliance on the presence of explicit slurs as a proxy for hate speech. This approach fails to account for the linguistic phenomenon of reclamation (or re-appropriation), where derogatory terms are repurposed by the target group, particularly the LGBTQ+ community, as markers of self-identification, pride, and solidarity [2].

In accordance with the previous EVALITA campaigns, in which both generic Hate Speech Detection (HaSpeeDe) [3] and Homotransphobia Detection (HODI) [4] challenges have been proposed in the context of the Italian language, the MultiPRIDE challenge [1] aims to explore the recognizing of usage of such slurs as reclamation terms and not as offensive one. The challenge is organized in two binary tasks: tasks A is a Text-based Detection, while task B is Text-based Detection with Context. The corpus consists of a multilingual dataset comprising tweets in three languages and the proposed subtasks are categorized according to linguistic affiliation: Italian (subtasks A1 and B1), Spanish (subtasks A2 and B2) and English (subtask A3). Data for task B adds the biography of the user (user bio) of the given tweet, if available.

This paper describes the system developed by the LlaNa team for subtasks A1 and B1 of the MultiPRIDE challenge in the context of the EVALITA 2026 campaign [5]. Our approach is grounded in a *data-centric* philosophy: rather than relying on complex ensemble architectures, we focused on addressing the

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.

alessio.mercurio@community.unipa.it (A. Mercurio); giorgio.talluto@community.unipa.it (G. Talluto); irene.siragusa02@unipa.it (I. Siragusa); roberto.pirrone@unipa.it (R. Pirrone)

0009-0005-8434-8729 (I. Siragusa); 0000-0001-9453-510X (R. Pirrone)

published 2026-04-16

primary bottleneck of the task, the scarcity of positive reclamation samples. We propose a robust fine-tuning pipeline based on XLM-RoBERTa [6], enhanced by a targeted cross-lingual data augmentation strategy. By translating reclamation instances from the auxiliary English and Spanish datasets into Italian, we successfully enriched the minority class, enabling the model to learn reclamation patterns more effectively.

The proposed system reached the 1st place in subasks B1 and reaches the 5th place in subtask A1, assessing the validity of the implemented approach.

This report is arranged as follows: Section 3 describes the proposed system along with data augmentation process, while obtained results and related discussion are reported in Section 4 and 5. Concluding remarks are in Section 6.

3. Description of the system

The MultiPRIDE dataset provided for subtasks A1 and B1 consists of 1,086 Italian annotated tweets for the presence of slurs used in a reclamatory manner. For the development phase, we performed a train-test split using an 80-20 ratio, resulting in 818 and 218 samples, respectively.

3.1. Data pre-processing

We adopted a conservative pre-processing strategy to preserve the linguistic nuances typical of social media communication. Although the provided dataset contained normalized tokens for user mentions @USER and URLs URL, we retained these tokens as they appear in the distribution, without removal. This decision diverges from standard approaches that strip such metadata: we hypothesize that interaction patterns (mentions) may implicitly correlate with the context of reclamation. Regarding textual content, we intentionally avoided lower-casing since it allows for a more accurate extraction of embeddings for case sensitive Language Models (LMs), such as XLM-RoBERTa [6] which is pre-trained on cased text. In addition, uppercase text conveys significant prosodic information and emotional intensity that a lower-cased normalization would destroy. Similarly, emojis were preserved in their original form, as they represent a fundamental component of online sentiment expression and may act as crucial contextual cues in the context of slur usage.

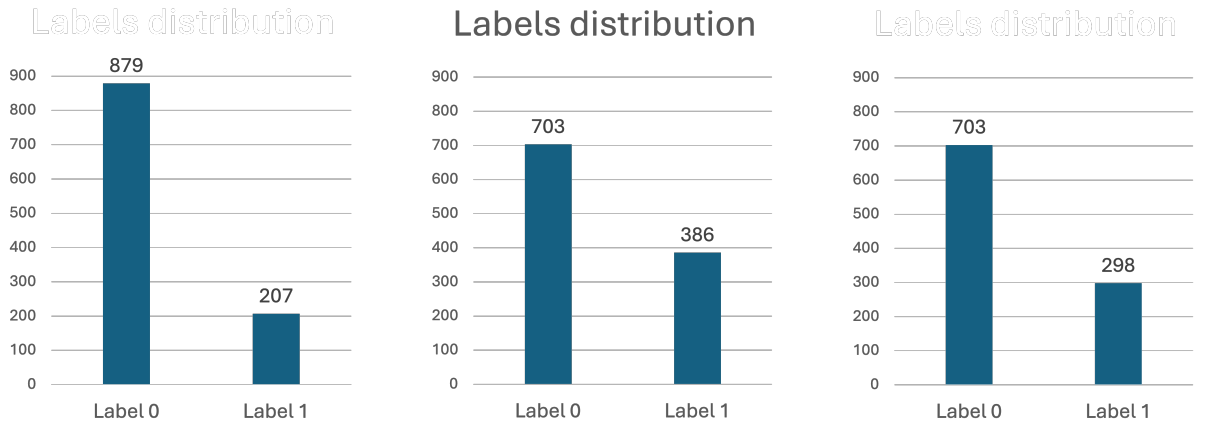

Figure 1: (a) Original unbalanced labels distribution in the Italian dataset (b) Labels distribution of the final dataset for subtask A1, using both English and Spanish dataset for augmentation (c) Labels distribution of the final dataset for subtask B1, using only the Spanish dataset for augmentation.

The primary challenge of the provided training set was the significant class imbalance (Figure 1a). The distribution featured a minority of positive samples (207 instances) compared to the negative ones (879 instances). Training deep neural networks on such skewed data typically leads to models that

overfit the majority class and fail to generalize on the phenomenon of interest. To address the data scarcity and class imbalance, we implemented a targeted cross-lingual augmentation pipeline [7], as shown in 2. Specifically, samples with a positive label, indicating the presence of reclamation, have been extracted from both English and Spanish datasets of subtasks A2, A3 and B2, and then have been machine-translated into Italian (Figure 1b). For subtask B1, the augmentation was restricted to the Spanish dataset (B2) (Figure 1c). This procedure let to the creation of two different augmented dataset for A1 and B1 tasks, where the latter had fewer positive samples to work with. Despite the resulted scarcisity, the developed model will take more advantage from the additional context provided by the user bio.

For the translation process, we selected the English-Italian and Spanish-Italian models from OPUS-MT [8, 9]. The selection of these specific open-source models, as opposed to larger generative models (LLMs), was a deliberate methodological choice, as we observed that the OPUS-MT models exhibit a lower tendency to censor or euphemize offensive terms. This characteristic was crucial for our task: it ensured that the translated tweets retained the explicit slurs necessary for the model to learn reclamation patterns, rather than having them replaced with neutral synonyms or removed entirely by a safety filter.

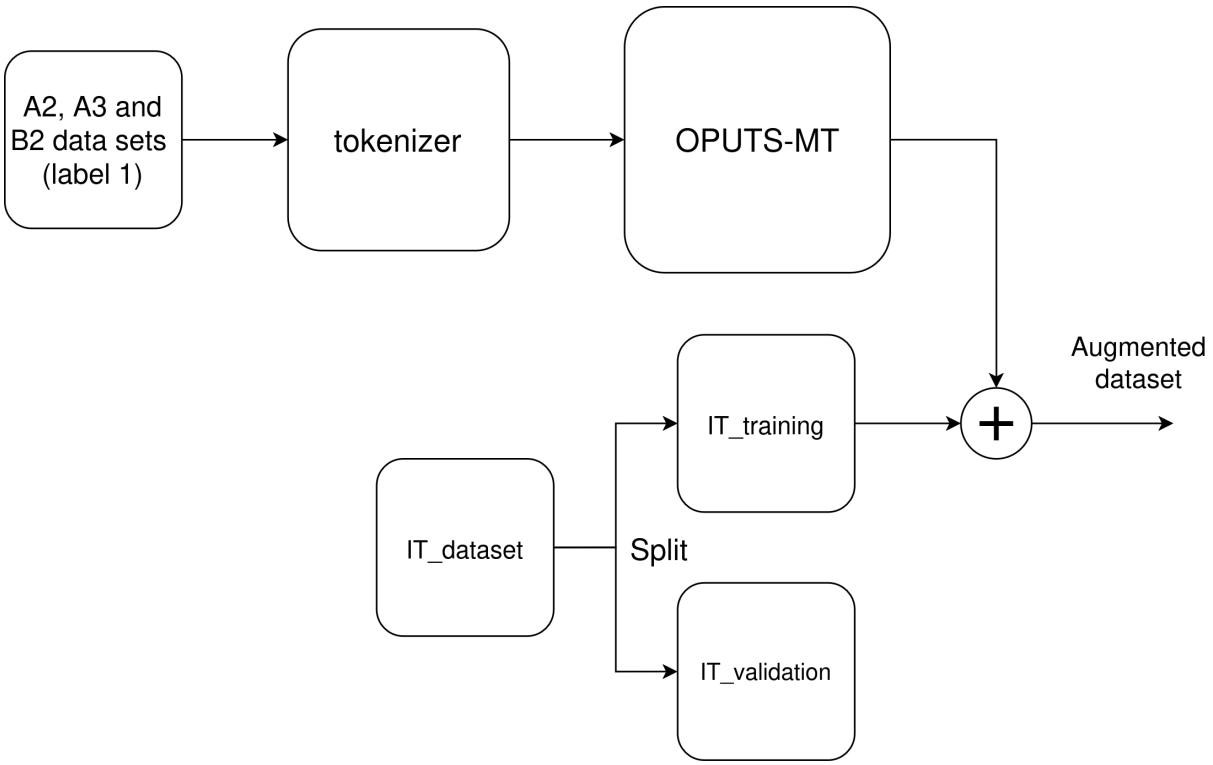

Figure 2: The proposed augmentation pipeline.

3.2. Network architecture

We employed a unique architecture for both tasks to evaluate whether the inclusion of context (user bio) yields performance improvements using the same developed model. The backbone of our system is XLM-RoBERTa-base [6, 10], selected for its state-of-the-art performance in multilingual settings and robustness on social media text. After the pre-processing phase, input sequences have been padded to a fixed length of 256 tokens. In the case of subtask B1, the input sequences are made up of the concatenation of the tweet and the user bio. The embedding of the CLS token is extracted from the last hidden layer of the encoder and fed into a custom classification head. This module consists of a dense layer projecting the 768-dimensional embedding to a hidden size of 512, followed by a normalization

layer and a ReLU activation function. To prevent overfitting, a dropout layer with a probability of 0.3 is applied before the final dense binary classifier. An overview of the developed architecture is reported in Figure 3.

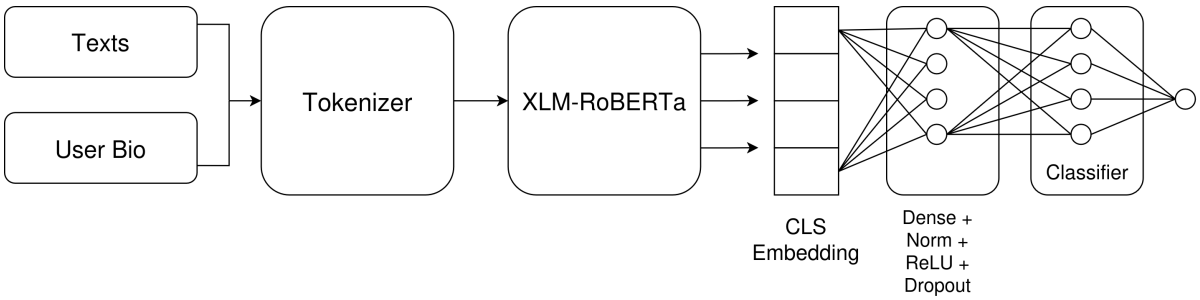

Figure 3: Schema of the proposed architecture made up of a module for embedding extraction, a dense layer with regularization followed by a binary classifier.

We trained both the LM and the dense layers on a workstation equipped with an AMD Ryzen 7 5800x CPU (32GB RAM) and an NVIDIA RTX 3060 GPU (12GB VRAM). To address the residual class imbalance, not fully resolved by the augmentation strategy, we computed and applied class weights within the loss function. The network has been optimized using AdamW [11] with a learning rate of 1e-5 and a weight decay of 0.05. We used a batch size of 8 and a weighted Binary Cross Entropy [12] as loss function. The fine-tuning process ran for a maximum of 8 epochs, employing an early stopping mechanism with a patience of 4 epochs.

4. Results

We conducted a two-phase evaluation to assess the effectiveness of our proposed pipeline. During the development phase, we evaluate the impact of the cross-lingual augmentation on our internal validation set, while in the evaluation phase, the best developed models have been tested with the official test set.

We evaluated the performance of the developed model with our internal validation set for both subtasks A1 and B1, and obtained results are reported in Table 1. More specifically, we both tested the impact of a simple training without data augmentation (w/o augmentation) and with the proposed data augmentation strategy (w/ augmentation).

Table 1
Performances over our internal validation set (macro F1-score) for both subtasks with and without data augmentation.

Strategy	Subtask A1	Subtask B1
w/o augmentation	0.8817	0.9033
w/ augmentation	**0.9131**	**0.9196**
Δ (w/ − w/o)	+0.0314	+0.0163

The experimental results demonstrate the effectiveness of the proposed data-centric strategy, and the implemented cross-lingual augmentation pipeline proved decisive for both subtracks. We observed a performance leap when training the models with the augmented data: the macro F1-score rose from 0.8817 to 0.9131 for subtask A1, and from 0.9033 to 0.9196 for subtask B1. This confirms that transferring reclamation patterns from English and Spanish to Italian helps the model generalize better on the minority class.

The comparison between subtask A1 and subtask B1 highlights the value of contextual feature. The inclusion of the user bio in subtask B1 resulted in a further improvement which is more prominent in the w/o augmentation pipeline, where an increment of 2.4% in the macro F1-score can be found. On the contrary, the increase in the w/ augmentation pipeline is marginal (+0.7%). This suggests that

the cross-lingual augmentation provides a robust enough representation of reclamation patterns to partially compensate for the absence of explicit user profiling. Consequently, considering the superior performance across both tasks, we selected the model trained with data augmentation as our model for the blind evaluation.

The official evaluation on the held-out test set confirmed the robustness of our system, which places above the baseline proposed by the author in both subtasks. In addition, the developed system reaches the 5th place for subtask A1, while winning the competition in subtask B1. The results from Tables 2 and 3 show an macro F1-score equal to 0.8835 on subtask A1, and to 0.9021 on subtask B1, which are consistent with those previously observed during the development phase. The highest performance on subtask B1 validates our hypothesis according to which extralinguistic context like the user bio is essential for disentangling reclamation instances, even on unseen data. More specifically, as well as the training split of the dataset, the only difference in the test set from subtasks A1 to B1 is the contextual field.

Table 2
Performances over the official test set for subtask A1. The official score for the final rank is reported in bold.

Class	Precision	Recall	F1-score
0	0.9451	0.9709	0.9579
1	0.8617	0.7625	0.8091
Macro	0.9034	0.8667	**0.8835**
Baseline			*0.8731*

Table 3
Performances over the official test set for subtask B1. The official score for the final rank is reported in bold.

Class	Precision	Recall	F1-score
0	0.9624	0.9624	0.9624
1	0.8417	0.8417	0.8417
Macro	0.9021	0.9021	**0.9021**
Baseline			*0.8981*

Analyzing the class-wise performance of the positive label, we achieved a precision of 0.86 but a lower recall of 0.76 in subtask A1 (Table 2). Conversely, in subtask B1, the model achieved a more balanced performance on the minority class (Table 2). This suggests that while textual features are sufficient for high-precision detection, the user profile context is essential for recalling more subtle instances of reclamation that might otherwise be ambiguous. The diverse class imbalance among the two subtasks reinforce this assumption: since for subtask B1 we had fewer samples for the data augmentation phase and the model had fewer samples to work with, the user bio resulted crucial for performance improvement.

Finally, the system exhibited high robustness on the negative class, consistently achieving a macro F1-scores above 0.95 across both subtasks.

5. Discussion

The obtained results highlights that the primary bottleneck of our model was the lack of a sufficient samples with reclamation, rather than the architecture itself. However, despite the encouraging results, the system is not infallible. As suggested by the organizers of the task, an error analysis was performed particularly on the tweets that were misclassified by the model.

A direct inspection of the misclassified tweets reveals that errors often stem from complex contexts where the boundary between offensive usage, reclamation, and mere reporting is blurred. A common

pattern involves tweets where the slur is embedded in complex sentence structures that mix conflicting sentiments (e.g. struggle vs. pride). Consider the following example:

> *"oscillo tra il volevo nascere etero e meno male che sono nato lgbt+ ed oggi più che mai grazie divinità egizie per avermi fatto r*cchione"*

In this case, the user expresses a journey from internalized conflict to self-acceptance, using slurs in a reclamatory manner. The model likely struggled to disentangle the initial negative sentiment *"volevo nascere etero"* from the final reclamation *"grazie... per avermi fatto..."*, misclassifying the intent.

A more subtle challenge is represented by tweets that contain slurs neither to offend nor to reclaim, but to report their usage by others.

> *"Fedez, che voi avete decretato paladino dei diritti civili, ha ripetuto per 3 volte le parole fr*cio e n***o nelle storie Instagram, complimenti"*

In this instance, slurs are used by the author to quote their usage from a public figure and the overall tweet is a critique towards him and has an ironic tone. The classifier incorrectly flagged this as reclamation because it detected the explicit terms without grasping the metalinguistic context.

6. Conclusion

In this paper we reported the architecture proposed by the LlaNa team for MultiPRIDE subtasks A1 and B1 promoted at the EVALITA 2026 campaign. The developed system won the competition for subtask B1 and reached the 5th place subtask A1. Our experimental results demonstrate that fine-tuning a multilingual encoder (XLM-RoBERTa), when combined with a targeted cross-lingual translation data augmentation strategy, achieves robust performance in correctly classifying the use of slurs in a reclamatory manner. In particular, the comparison between the two subtasks highlights the importance to leverage extralinguistic context. The usage of the user bio, as context, resulted in a more robust classification performance, particularly in validating the generalization capabilities of developed model on unseen data. In future works we will focus on the integration of more sophisticated translation models that preserve prosodic features to better overcome issues related to data scarcisity.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3 Pro in order to: Grammar and spelling check, Paraphrase and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

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] A. Galinsky, C. Wang, J. Whitson, E. Anicich, K. Hugenberg, G. Bodenhausen, The reappropriation of stigmatizing labels: The reciprocal relationship between power and self-labeling, Psychological science 24 (2013). doi:10.1177/0956797613482943.

[3] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, Overview of the EVALITA 2018 Hate Speech Detection Task, in: Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2018), CEUR-WS.org, 2018.

[4] D. Nozza, E. Fersini, V. Patti, F. Rangel, P. Rosso, HODI at EVALITA 2023: Overview of the Homotransphobia Detection Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2023), CEUR-WS.org, Parma, Italy, 2023.

[5] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, 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.

[6] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, et al., Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[7] T. Ranasinghe, M. Zampieri, Multilingual offensive language identification with cross-lingual embeddings, 2020. URL: https://arxiv.org/abs/2010.05324. arXiv:2010.05324.

[8] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato, Y. Scherrer, R. Vazquez, S. Virpioja, Democratizing neural machine translation with OPUS-MT, Language Resources and Evaluation (2023) 713–755. doi:10.1007/s10579-023-09704-w.

[9] J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the World, in: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.

[10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al., Huggingface's transformers: State-of-the-art natural language processing, 2020. URL: https://arxiv.org/abs/1910.03771. arXiv:1910.03771.

[11] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. URL: https://arxiv.org/abs/1711.05101. arXiv:1711.05101.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., Pytorch: An imperative style, high-performance deep learning library, 2019. URL: https://arxiv.org/abs/1912.01703. arXiv:1912.01703.