GRUPPETTOZZO at MultiPRIDE 2026: Detecting LGBTQ+ Reclamatory Intent via Context-Aware Transformers

Federico Traina[1,†], Alessandro Santoro[1,†], Gabriele Greco[1,†], Irene Siragusa[1,*] and Roberto Pirrone[1]

[1]*Department of Engineering, University of Palermo, Palermo, 90128, Sicily, Italy*

Abstract

In this report we presents the proposed approach of GRUPPETTOZZO team for the detection of reclamatory usage of terms related to LGBTQ+. In the context of the multilingual MultiPRIDE challenge at EVALITA 2026 campaign, we proposed a encoder-only transformer-based method to classify the textual content of the given tweets (task A) and in conjunction with the contextual information provided by the user biography (task B). To handle the unbalance in the given data set, we explored the usage of both weighted and focal loss functions and two data augmentation strategies, to better generalize the performance of developed model in the target multilingual context. The best developed models achieved macro F1-scores exceeding 0.88 for Italian and 0.73 for Spanish-based tasks, while reaching 0.59 for the English based one.

Keywords

Context-Aware Transformers, LGBTQ+ Slurs Reclamation, Data Augmentation, Language Models

1. Introduction

Language within the LGBTQ+ context is characterized by complex phenomena such as reclamation, whereby terms originally intended as slurs are repurposed by community members for positive and/or identity-affirming purposes. In the context of social media and hate speech detection applications, an important challenge lies in distinguishing the correct usage of such slurs which may have both a reclamation or a denigratory intent. MultiPRIDE challenge [1] attempts to address this complex phenomenon proposing a two multilingual binary classification tasks. In particular, the objective is to determine wherever a given tweet, uses a term related to LGBTQ+ context with a reclamatory intent (task A) in Italian (subtask A1), Spanish (subtask A2) and English (subtask A3). In addition, organizers add the user biography (user bio) when available as extra context for the classification (task B) in Italian (subtask B1) and Spanish (subtask B2), as well as encouraging cross-lingual applications (subtask A-multi and B-multi).

This paper describes the system developed by the GRUPPETTOZZO team for tasks A and B of the MultiPRIDE challenge in the context of the EVALITA 2026 campaign [2]. Our approach consist in fine-tuning encored-only Language Models (LMs) to address task A and, for task B, we integrated also textual biographies as pragmatic context to model's input. Moreover, we handle tasks A and B in both monolingual and multilingual configuration. Our approach can be summarized in five stages, as can be seen in Figure 1. Firstly we perform a data augmentation phase to enrich the original data set, following two different strategies. Textual content are pre-processed through a data cleaning procedure before being injected into the developed Context-Aware Architecture, made up of a LM backbone, a custom pooling strategy and a classification head that performs the the final prediction.

This report is structured as follows: data augmentation and pre-processing procedures, and 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.

federico.traina01@community.unipa.it (F. Traina); alessandro.santoro04@community.unipa.it (A. Santoro); gabriele.greco07@community.unipa.it (G. Greco); irene.siragusa02@unipa.it (I. Siragusa); roberto.pirrone@unipa.it (R. Pirrone)

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

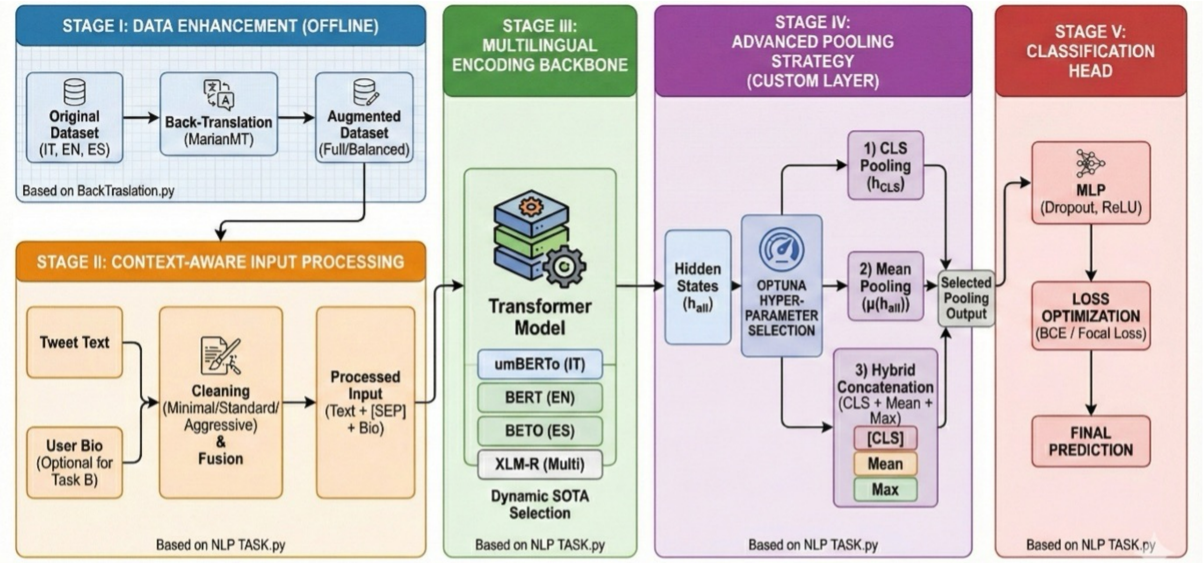
STAGE I: DATA ENHANCEMENT (OFFLINE)
Original Dataset (IT, EN, ES)
Back-Translation (MarianMT)
Augmented Dataset (Full/Balanced)
Based on BackTraslation.py
STAGE II: CONTEXT-AWARE INPUT PROCESSING
Tweet Text
User Bio (Optional for Task B)
Cleaning (Minimal/Standard/ Aggressive) & Fusion
Processed Input (Text + [SEP] + Bio)
Based on NLP TASK.py
STAGE III: MULTILINGUAL ENCODING BACKBONE
Transformer Model
umBERTo (IT)
BERT (EN)
BETO (ES)
XLM-R (Multi)
Dynamic SOTA Selection
Based on NLP TASK.py
STAGE IV: ADVANCED POOLING STRATEGY (CUSTOM LAYER)
Hidden States (h_{all})
OPTUNA HYPER-PARAMETER SELECTION
1) CLS Pooling (h_{CLS})
2) Mean Pooling ($\mu(h_{all})$)
3) Hybrid Concatenation (CLS + Mean + Max)
[CLS]
Mean
Max
Selected Pooling Output
Based on NLP TASK.py
STAGE V: CLASSIFICATION HEAD
MLP (Dropout, ReLU)
LOSS OPTIMIZATION (BCE / Focal Loss)
FINAL PREDICTION
Based on NLP TASK.py

Figure 1: Overview of the proposed approach.

proposed system are described in Section 2, while results and related discussion are presented in Sections 3 and 4. Section 5 contains final remarks.

2. Description of the system

MultiPRIDE data set is arranged in three different splits, one for each language, comprehensive of 1,000 labeled tweets, where positive classes are associated to usage of slurs in a reclamatory manner. For Spanish and Italian splits, the user bio field was provided to address task B, if available. For the development phase, we split the data set using an 80-20 ratio and a stratified strategy.

The system processes data in two distinct structural modalities designed to evaluate the impact of external context on classification, depending exclusively on the target task and not on the language:

- **Text-Only** the input to the model is exclusively the textual content of the tweet (subtasks A1, A2, A3, A-multi);
- **Context-Aware** an enriched representation is generated by concatenating the original tweet with the user bio (subtasks B1, B2, B-multi). The input follows the pattern:

  ```
  TWEET [SEP] CONTESTO: BIO
  ```

  where the `[SEP]` token serves as a segment delimiter, thus enabling the model to distinguish between the two distinct information sources within the same input sequence. Whenever the user bio is not available, it is set to an empty string.

2.1. Data augmentation and pre-processing

Since the unbalanced distribution of positive labels and the limited number of samples in the training set, we implemented a back-translation pipeline based on MarianMT [3] and OPUS-MT [4, 5] models. This technique involves translating the text from the source language to a pivot language and subsequently re-translating it back into the original language. This process generates a paraphrase that preserves the core semantic content while introducing lexical and syntactic variance. We used English as pivot language for Italian and Spanish splits, while Spanish as pivot language for English split.

We experimented two augmentation configurations as shown in Figure 2:

- **Full Augmentation** A paraphrase is generated for *every sample* in the dataset. On doing this, we double the training set size, exposing the model to greater linguistic variability and helping to prevent overfitting on specific syntactic patterns.
- **Balanced Augmentation** Augmentation is applied *exclusively* to the minority class, i.e. the positive labeled samples. This technique reduces the class imbalance ratio, enabling the model to learn more robust features for the positive class without resorting to naive oversampling, which would simply duplicate existing examples.

After the data augmentation phases, a check over the generated sample has been conducted and eventually duplicated samples, which may result if the back-translation system outputs the original text, have been removed.

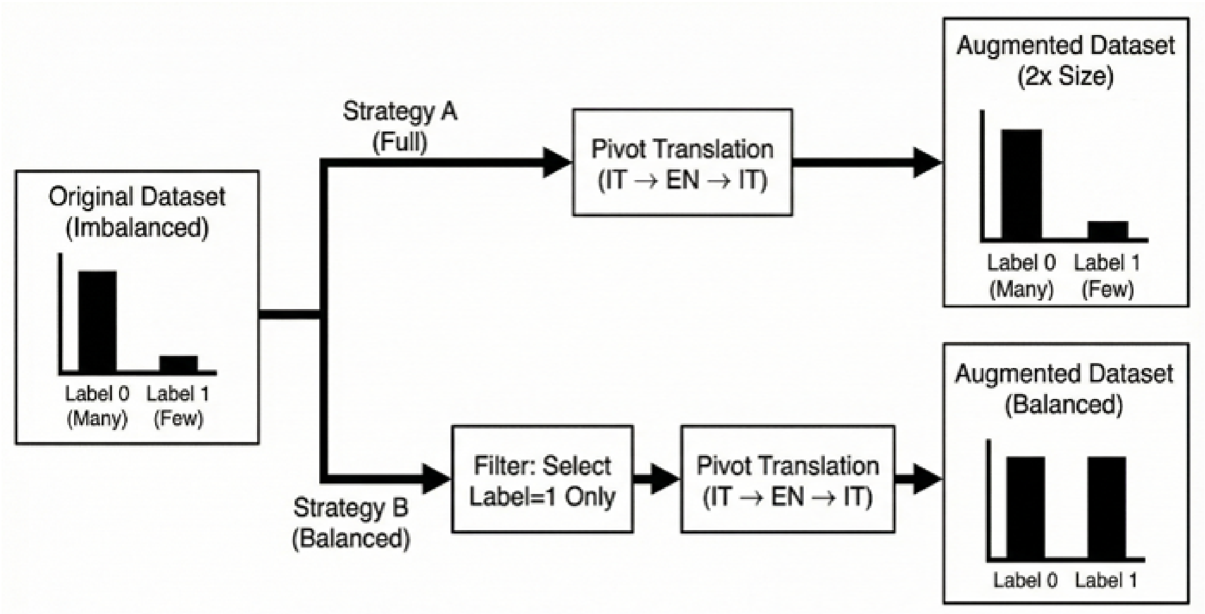

Figure 2: Overview of the data augmentation pipeline with full and balanced augmentation strategies.

Given the social media nature of the data set, we formalized three different strategies of data preprocessing, and we investigated the impact of each strategy:

- **Minimal** the original text is converted in lowercase and no further data cleaning operation are performed, thus the textual content is left unchanged and are kept both the total amount of information and any noisy information.
- **Standard** a demojization phase is performed over the lowercased input text, and any present emoji is converted in a textual string, e.g. 🏳️‍🌈 ⟶`:rainbow_flag:`.
- **Aggressive** both emoji and stopwords are removed and textual content is lowercased.

The choice to lowercase the input in all the pre-processing strategies is a strategic decision for our task. On social media, capitalization is often irregular or used exclusively for emphasis. Normalizing the text helps the model focus on semantic content rather than graphical variations, thus enhancing its generalization capabilities.

Emojis are dense semantic and sentiment information. Removing them eliminates any potential signals of LGBTQ+ community affiliation (e.g., colored hearts, pride flags) or references to the sentiment of the given tweet. The demojization projects these symbols into the semantic space of the BERT tokenizer [6], allowing the model to proper interpret those meta-linguist features. In the aggressive strategy, the stopwords removal can be detrimental in pragmatic tasks. Functional particles, such as negations or pronouns, which can radically alter the sentence's meaning.

2.2. Context-Aware Architecture

The proposed Context-Aware architecture departs from standard classification models by integrating modular components designed to maximize semantic feature extraction and enhance model robustness in the presence of imbalanced data. It is made up of a LM backbone, followed by a neural module which further processes the dense representation obtained from the LM with an custom pooling strategy and a classification head that performs the the final prediction.

2.2.1. Language Model backbones

The proposed model leverages on fine-tuning of pre-trained transformer models [7], selected to maximize the semantic representation in each target language.

For the Italian split, we used lupobricco/umBERTo_fine-tuned_hate_offensivity[1] (umBERTo). This model has been obtained after a domain-specific fine-tuning of UmBERTo [8], which is an Italian LM based on RoBERTa model [9], and trained over a vast Italian corpora using the Whole Word Masking (WWM) technique. Specifically, the WWM requires the model to predict entire words even when only a sub-word unit is masked. This results to be a crucial feature in the context of Romance languages (Italian, Spanish), which are characterized by an abundance of suffixes and complex conjugations. In addition, UmBERTo more effectively handles the complex morphology of the Italian language by utilizing a SentencePiece-based tokenizer, which better preserves the semantic roots of words. The choice of umBERTo provides a critical advantage in Task A. This specific version has been further fine-tuned for Hate Speech Detection, thus resulting (i) already optimized to recognize patterns of toxicity and (ii) the need of an extensive training from scratch to identify aggressive intent is significantly reduced.

For the English split we adopted the standard BERT model in its base and uncased version [6], which represent a state-of-the-art LM.

As for the Spanish data set, we used BETO [10], a BERT-based model trained with the WWM technique over Spanish data. Analogously for the chosen Italian LM, WWM allow a better modeling of the Spanish language and the model itself is forced to learn deeper syntactic and semantic representations, consistently outperforming standard Multilingual BERT (mBERT) [6].

We also considered a multilingual backbone for the cross-lingual classification task. We opt for XLM-RoBERTa [11], which scales cross-lingual pre-training utilizing 2.5TB of filtered CommonCrawl text across 100 languages. Unlike mBERT, XLM-RoBERTa omits the Next Sentence Prediction (NSP) objective and is trained with significantly larger batches and longer durations. This architecture facilitates the creation of a shared semantic space where semantically equivalent terms in different languages (e.g. 'gay' in English and 'gay' in Italian) are mapped to proximal vectors. This cross-lingual alignment capability is fundamental to our system, enabling the transfer of knowledge from high-resource examples in English and Spanish to improve predictions even for underrepresented classes [12]

2.2.2. Neural module architecture

Despite the different LM backbone used, the output of the last layer $H \in \mathbb{R}^{l \times d}$, where l is the length of the given sequence and d the dimension of the hidden state, serves as input to a neural model.

From this vector representation, three pooling strategies are defined:

- **CLS Pooling** $\mathbf{v}_{cls}$ which selects the embedding of the `[CLS]` token;
- **Mean Pooling** $\mathbf{v}_{mean}$ across the target sentence removing any padding tokens;
- **Hybrid Pooling** is the proposed advanced concatenation strategy, which consist in concatenating the vector representation obtained by the previous pooling strategies and a Max Pooling $\mathbf{v}_{max}$ over the dense representations for each sample removing any padding token. The final representation is obtained as $\mathbf{v}_{final} = \mathbf{v}_{cls} \oplus \mathbf{v}_{mean} \oplus \mathbf{v}_{max}$, with dimension equal to $3 \times d = 2304$.

[1] https://huggingface.co/lupobricco/umBERTo_fine-tuned_hate_offensivity

While $\mathbf{v}_{cls}$ is optimized during LM training, $\mathbf{v}_{mean}$ captures the global context, and $\mathbf{v}_{max}$ highlights salient signals (e.g., a single offensive term). The concatenation of these vectors provides a more comprehensive overview of the input data to the final classification.

The classification head applies, over the obtained data after pooling, a dropout layer to prevent the overfitting of the network, followed by a hidden dense layer and a ReLU activation function and the classification layer.

2.3. Training Details

For the training phase, we adopted an hyperparameter optimization strategy, including architectural details such as the pooling strategy, dropout value, the size of the hidden MLP layer, and training details as the number of layers to freeze in the LM backbone, loss function, optimizer and learning rate, described below.

To balance learning capacity with the stability of pre-trained weights in the LM backbones, we implemented a selective freezing mechanism. The initial embeddings and the firsts n encoder layers can be frozen in order to (i) force the model to leverage its acquired foundational linguistic knowledge and (ii) focus the training process exclusively on the higher-level layers. Such high-level features, then combined with the neural model architecture, can deeply focusing toward the target task during its training phase.

We considered two different options as for the loss function: a weighted Binary Cross-Entropy (BCE) loss with the inverse class ratio as weight, and a Focal Loss [13], parametrized with γ, optimized to down-weight easy samples, and compelling the model to focus on more complex, hard-to-classify instances.

As optimizer during the training phase, we both evaluated AdamW [14] and EvoLved Sign Momentum (Lion) [15]. Unlike AdamW, which computes the first and second moments (mean and variance) of the gradients, the Lion optimizer utilizes only the sign of the gradient for weight updates. This approach simplifies the optimization process, reduces memory overhead, and provides a more robust regularization effect, specifically in fine-tuning on small data sets such as the MultiPRIDE one.

We used Optuna [16], a bayesian optimization to determine the optimal hyperparameter configuration for each task and subtask, and a comprehensive overview of all the considered hyperparameters reported Table 1.

Table 1
Overview of all the considered hyperparameters.

Hyperparameter	**Search space / possible values**
Pooling strategy	CLS Pooling, Mean Pooling, Hybrid Pooling
# frozen layers LM	1,2,3,4,5,6
MLP size	128, 256, 623
Dropout	from 0.1 to 0.5
Loss Function	Weighted BCE, Focal Loss
Gamma (Focal Loss)	from 0.5 to 3
Learning Rate	from 1e-5 to 5e-5
Optimizer	AdamW, Lion

The developed system and the training pipeline have been implemented in PyTorch [17] and with HuggingFace Transformers [18]. To optimize the computational efficiency we adopted a Mixed Precision Training (FP16) technique. We trained the models up to a maximum of 8 epochs and implemented an Early Stopping strategy over the macro F1-score in our validation set with a patience value equal to 3. Experiments have been conducted over Google Colab instances equipped with one GPU Tesla T4 and Kaggle instances with one NVIDIA TESLA P100 GPU.

3. Results

For each subtask diverse training configuration have been explored, namely the diverse augmentation strategies (without, full and balanced), data cleaning techniques (minimal, standard, aggressive), combined with the best hyperparameter obtained with OPTUNA for each configuration (Table 1). In this section we report the configurations which obtained the best results in terms of macro F1-score over our validation set and the official test set for both tasks.

In Table 2 are reported the obtained results for Task A, within the different augmentation strategies for each language and with the multilingual configuration.

Table 2
Results over validation set for Task A. Bold results refer to the highest ones per language split.

Subtask	Data Cleaning	Data Augmentation	Pooling	# frozen LM layers	MLP size	Optimizer	Loss	Macro F1-score
A1	Standard	None	CLS	2	128	Lion	BCE	0.9014
A2	Minimal	None	Mean	1	128	AdamW	BCE	0.7320
A3	Minimal	None	CLS	4	128	Lion	Focal	0.6288
A-Multi	Aggressive	None	Hybrid	5	128	AdamW	Focal	0.7898
A1	**Minimal**	**Full**	**Hybrid**	**4**	**128**	**Lion**	**BCE**	**0.9702**
A2	**Minimal**	**Full**	**CLS**	**6**	**128**	**Lion**	**Focal**	**0.9047**
A3	Minimal	Full	Mean	4	128	AdamW	Focal	0.8849
A-Multi	Standard	Full	Mean	3	128	Lion	Focal	0.8816
A1	Minimal	Balanced	Hybrid	0	128	Lion	Focal	0.9478
A2	Minimal	Balanced	Mean	2	128	Lion	Focal	0.8987
A3	**Standard**	**Balanced**	**Hybrid**	**4**	**256**	**Lion**	**Focal**	**0.9190**
A-Multi	**Standard**	**Balanced**	**Mean**	**5**	**128**	**Lion**	**Focal**	**0.8871**

Lowest results are obtained without data augmentation, suggesting an intrinsic complexity of the task, especially in contest of data scarcisity. Both strategies of full and balanced data augmentation led to higher performances, almost over 0.90 for all configurations, confirming that the initial poor generalization capabilities are mainly addressed to the the reduced number of training samples.

The Italian split (subtask A1) shows the overall best results, even without data augmentation and increasing with both full and balanced augmentation. This behavior may suggest an a minor linguistic ambiguity in Italian tweets, when compared with other languages. While both Italian and Spanish data benefits from the full augmentation strategy, English and multilingual splits highly benefit of the balanced augmentation.

As for data cleaning, the minimal strategy, which essentially left the input data unchanged, resulted to be the overall best choice, thus assessing that punctuation elements and emojis are semantically essential for the given task. Multilingual setup, on the contrary, needs a standard or aggressing data pre-processing strategy. These performances show that a normalization step helps cross-lingual models to better map concepts in a common latent and embedding space.

Regarding the pooling strategies, results obtained with the concatenation of the three pooling operation (hybrid pooling) results the best one for Italian and English splits, while Spanish split benefits from the usage of the simple `[CLS]` token, while the mean pooling was preferred for the multilingual setup.

The conjunct use of the focal loss with Lion optimizer resulted overall the best choice to focus the learning towards the more complex samples and the unbalanced class. Furthermore, the optimal size for the MLP layer was equal to 128 and overall best results are obtained though a soft training of the LM backbone, that is with an higher lever of transformer layers kept frozen during the training.

In Table 4 are reported the obtained results for Task B, within the different augmentation strategies for each language and with the multilingual configuration. In this task, the additional information provided by the user bio, introduces identity elements about the author of the target tweet, which can used by the model as suggestion for the final classification.

Table 3
Results over validation set for Task B. Bold results refer to the highest ones per language split.

Subtask	Data Cleaning	Data Augmentation	Pooling	# frozen LM layers	MLP size	Optimizer	Loss	Macro F1-score
B1	Aggressive	None	CLS	1	256	AdamW	Focal	0.9043
B2	Standard	None	Hybrid	1	128	AdamW	Focal	0.7064
B-Multi	Aggressive	None	Hybrid	1	256	AdamW	Focal	0.8340
B1	**Standard**	**Full**	**Mean**	**4**	**128**	**AdamW**	**Focal**	**0.9667**
B2	**Standard**	**Full**	**CLS**	**0**	**256**	**AdamW**	**Focal**	**0.9541**
B-Multi	Standard	Full	Hybrid	3	128	Lion	BCE	0.9015
B1	Standard	Balanced	Hybrid	3	128	Lion	BCE	0.9602
B2	Aggressive	Balanced	Hybrid	5	128	AdamW	Focal	0.9010
B-Multi	**Standard**	**Balanced**	**Hybrid**	**0**	**128**	**Lion**	**Focal**	**0.9041**

Also for this case, the full augmentation results the best data augmentation option, with the only exception for the multilingual case, which slightly benefits from the balanced augmentation. Overall, performances of the developed models without data augmentation are considerably lower compared to the ones with data augmentation.

Despite the additional information provided by the user bio, this did not result sufficient to compensate both the unbalanced labels distribution and the poor quantity of samples, making them comparable to the results obtained with the task A. More specifically, Italian and Multilingual splits benefits from the additional information of the user bio, compared to the Spanish one, whose performances decrease. We hypothesize that the model can effectively benefits of this additional context for the cases in which a high coherence is present between the user bio and the target tweet, suggesting that in the Italian split this correlation is higher than in the Spanish one. However, the overall best performances are obtained with the full data augmentation strategy and implementing the standard data cleaning method and a training based on Focal Loss and AdamW as optimizer.

For each task, we submitted the prediction obtained with the model reaching the best performances over our validation set (Tables 2 and 4) and in Table **??** are collected the obtained results over the test set.

Table 4
Results over validation set for Task B. Bold results refer to the highest ones per language split.

Subtask	Macro Precision	Macro Recall	Macro F1-Score
A1	0.914	0.859	0.883
A2	0.727	0.788	0.751
A3	0.587	0.618	0.598
B1	0.906	0.890	0.898
B2	0.729	0.731	0.730

Obtained results confirms the excellent performance of the fine-tuned umBERTo model. Specifically, in subtask B1 we achieved a macro F1-score of 0.898. The integration of user biographies yielded a +1.7% improvement over subask A1, confirming that biographical context serves as a robust field of information for disambiguating intent in the Italian dataset.

A drastic performance decline was observed for the English split in subtask A3, compared to the result observed during the development phase, where we reached an macro F1-score of only 0.598. This suggests either significant overfitting on the training data or a semantic distributional shift in the test set compared with the training one. The different nature of the data may led the model to struggled to differentiate between reclaimed usage and actual hate speech.

For Spanish subtasks, the system maintained consistent performance reaching 0.751 and 0.730 as macro F1-score in subtask A2 and B2 respectively. In contrast to the results observed for the Italian language, a significant improvement of +2.9% is found in subtask A2 over B2. This indicates that for the Spanish test set, biographical information may have introduced contextual noise rather than clarifying the intent of the speaker.

4. Discussion

A deeper analysis over the obtained results allow us to better analyze the complexity of the task and compare the effectiveness of the proposed techniques.

A counterintuitive finding from our experiments is the superiority of less invasive data cleaning strategies (minimal and standard) compared to the aggressive one. In traditional NLP tasks, stopwords and punctuation are often discarded as noise. However, within the LGBTQ+ linguistic context, the use of emojis (e.g., 🏳️🌈, ✨) and pronouns frequently serves as a crucial indicator of the author's identity and its possible reclaimed intent. The removal of these elements, in the aggressive strategy, deprives the model of fundamental pragmatic signals.

The implementation of a augmentation strategy with Back-Translation proved to be pivotal. Models trained exclusively on the original dataset exhibited clear signs of data scarcity and struggled with generalization. The introduction of augmented variants stabilized the training process, enabling the multilingual model (XLM-RoBERTa) to achieve performance levels comparable to those of monolingual models, reaching a macro F1-score higher than 0.90 in development phase in tasks B. This demonstrates a significant cross-lingual transfer learning capability of the model, where the model successfully leverages augmented patterns across the three languages.

An outperforming pooling strategy cannot be uniquely identified, but the overall hightest performance in both development and testing phases are addressed to the proposed hybrid pooling. This suggests that the salient information is not merely concentrated in the *[CLS]* token of the sequence, but is distributed across the tweet and the biography (mean and max pooling).

Moreover, the usage of the focal loss during the training guaranteed a more robust and generalization capabilities in the model itself which performances did not deteriorated significantly during the testing phase, with the only exception for subtask A3.

5. Conclusion

In this paper we reported the architecture proposed by the GRUPPETTOZZO team for MultiPRIDE tasks A and B promoted at the EVALITA 2026 campaign. Our experimental results demonstrate that the identification of reclamation intent in slurs' usage is not a mere lexical problem, but a semantic one. The best developed models achieved macro F1-scores exceeding 0.88 for Italian and 0.73 for Spanish-based tasks, while reaching 0.59 for the English based one. Obtained results shows that preserving the non-verbal pragmatic markers (emojis) while adopting a suitable data augmentation strategy and a training procedure based on a focal loss, were our core features to increment model performance over the two proposed tasks. The concatenation of the textual content and the user bio for tasks B, allowed a performance increment when compared to tasks A for the Italian split, while the opposite behavior is found for the Spanish one.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3 in order to: Text Translation, Grammar and spelling check, Paraphrase and reword. Further, the author(s) used Gemini 3 for figures 1, 2 and 3 in order to: Generate images. 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] L. Draetta, C. Ferrando, 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 Reclamation of Slurs 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] 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.

[3] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, et al., Marian: Fast neural machine translation in C++, in: Proceedings of ACL 2018, System Demonstrations, Melbourne, Australia, 2018, pp. 116–121. URL: https://aclanthology.org/P18-4020/. doi:10.18653/v1/P18-4020.

[4] 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.

[5] 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.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, volume 1 (long and short papers), 2019, pp. 4171–4186.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[8] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, GitHub, 2020.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[10] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[11] 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.

[12] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 907–914. URL: https://aclanthology.org/2021.acl-short.114/. doi:10.18653/v1/2021.acl-short.114.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007. doi:10.1109/ICCV.2017.324.

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

[15] X. Chen, C. Liang, D. Huang, E. Real, K. Wang, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, Y. Lu, Q. V. Le, Symbolic discovery of optimization algorithms, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2023.

[16] 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, New York, NY, USA, 2019, p. 2623–2631. URL: https://doi.org/10.1145/3292500.3330701. doi:10.1145/3292500.3330701.

[17] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: https://proceedings.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.