# UniBO-FICLIT at MultiPRIDE: Fine-Tuning an ELECTRA-Based Model for the Detection of Italian Reclaimed Slurs

Simone Casazza[1,*]

[1]University of Bologna, Bologna, Italy

**Abstract**

Linguistic reclamation, where targeted communities reappropriate slurs or identity-related terms for non derogatory, affiliative, or dissociative purposes, poses a distinctive challenge for hate speech detection in NLP. This report presents an encoder-based system for the EVALITA 2026 MultiPRIDE task on Italian tweets, employing an ELECTRA architecture fine tuned for binary sequence classification in two settings: text only and text plus limited author context. Results indicate that the system reliably recognizes tweets that do not contain reclaimed language, whereas detecting tweets that do contain reclamation remains difficult, even when author biographies are included. Error analysis highlights three recurrent drivers of misclassification: orthographic variation in slurs, explicit attributive uses by users, and ambiguous stance signals in the absence of clear context. These findings underscore the methodological limits of a strictly binary framing. The study motivates broader modeling perspectives alongside practical steps within current pipelines, including more robust preprocessing to mitigate orthographic noise and exploration of data augmentation and rebalancing strategies under skewed training distributions.

**Keywords**

ELECTRA, fine-tuning, linguistic reclamation

***Warning**: This paper contains examples of explicitly offensive content.*

## 1. Introduction

Hate speech detection is a well-established task in natural language processing (NLP), yet comparatively little attention has been devoted to the closely related phenomenon of linguistic reclamation, the practice whereby targeted communities reappropriate slurs or identity-related terms for non-derogatory, affiliative, or dissociative purposes. From a computational perspective, this poses a distinctive challenge: the same lexical item may function either as an expression of hostility or as a marker of in-group solidarity, irony, or critical stance, depending on context, speaker identity, and audience uptake. In operational systems, this ambiguity can inflate false positives, which not only degrade model utility but risk further marginalizing LGBTQ+ users by erroneously flagging and suppressing their speech, thereby reproducing harms against a group that is already subject to social and institutional oppression.

This report investigates reclamation detection within the MultiPRIDE framework [1] at EVALITA 2026, focusing on Italian tweets that contain LGBTQ+-related terms. The approach adopts an encoder-based ELECTRA architecture fine-tuned for binary sequence classification, examining a text-only setting alongside a setting that incorporates limited author context. The analysis highlights two structural hurdles, class imbalance and the pragmatic complexity of reclamation, that jointly hinder reliable identification. Results indicate that recognizing tweets without reclaimed language is comparatively easier than identifying reclamation itself, underscoring the limits of a binary framing for a context-dependent, multi-dimensional phenomenon.

Beyond establishing a baseline through fine-tuning ELECTRA, the study maps error concentrations

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

[*] Corresponding author.

simone.casazza@studio.unibo.it (S. Casazza)
https://github.com/simocasaz (S. Casazza)
00009-0003-2593-9030 (S. Casazza)

around orthographic variation, attributive uses, and ambiguous stance cues. Taken together, the findings motivate a broader modeling perspective beyond strict binary outcomes, alongside practical steps, more robust preprocessing to handle orthographic noise and systematic exploration of data augmentation and rebalancing, to strengthen learning under skewed and limited data conditions.

## 2. Related work

Hate speech detection as a natural language processing (NLP) task has been widely addressed in the literature (Fortuna and Nunes 2018; Poletto et al. 2021) [2] [3], but little attention has been given to the phenomenon of linguistic reclamation. This phenomenon has long been studied by linguists and philosophers of language (Brontsema 2004; Hom 2008; Anderson and Lepore 2013; Bianchi 2014; Cepollaro and López de Sa 2023) [4] [5] [6] [7] [8] because of its close connection to the nature of slurs and, consequently, to the nature of meaning in language.

From an NLP perspective, recent work has begun to address this challenge. Zsisku et al. (2024) [9] highlight a pragmatic issue: reclaimed language used by targeted groups can be misclassified as hate speech by automatic detectors that are not trained to recognize reclamation. To mitigate this, the authors introduced the RHDS dataset and fine-tuned a RoBERTa model previously trained on hate speech detection. Their approach significantly reduced false positives.

For Italian, the only relevant study is by Cuccarini et al. (2024) [10], who adapted the HODI dataset (Nozza et al. 2023) [11] by filtering tweets containing slurs using 17 terms from the HurtLex lexicon (Bassignana et al. 2018) [12]. They employed Qwen, a multilingual decoder-based model pre-trained on Italian, in a zero-shot setting with four progressively specific prompts.

Generally speaking, fine-tuning pre-trained encoder models for text classification tasks, especially in hate speech detection, has proven effective. For instance, Lavergne et al. (2020) [13] developed the best-performing model in the EVALITA 2020 HaSpeeDe 2 task (Sanguinetti et al. 2020) [14], while Locatelli and Locatelli (2023) [15] achieved top results in the EVALITA 2023 HODI task by fine-tuning a transformer-based model. Both cases confirm the strong performance of encoder architectures in Italian hate speech detection.

## 3. Task description

The MultiPRIDE task at EVALITA 2026 addresses the challenge of detecting reclaimed language in the LGBTQ+ context, where slurs or identity-related terms are used with a reappropriative intent rather than as hate speech. Participants are asked to build systems capable of performing binary classification: given a sentence containing an LGBTQ+-related term, the system must determine whether the term is used in a reclamatory way or not.

The task is organized into two main settings. The first focuses exclusively on the textual content of the message, allowing participants to work in either a constrained mode, using only the provided training data, or an unconstrained mode, where additional annotated resources may be incorporated. The second setting extends the classification to include contextual information, specifically the author's profile biography when available, alongside the text. Both settings include subtasks for different languages, with Italian, Spanish, and English covered in the textual setting, and Italian and Spanish in the contextual setting.

Evaluation is based on macro F1-score computed over the binary label. Baseline systems are provided by fine-tuning language-specific BERT models with weighted cross-entropy loss.

## 4. Dataset

The dataset for the Italian subtasks, sourced from the TWITA collection (Basile et al. 2018) [16], has been divided by the organizers into two distinct sets: a training set that constitutes 60% of the entire

**Table 1**
Training set label distribution

| Label | Frequency |
|---|---|
| 1 (reclamation) | 879 |
| 0 (no reclamation) | 207 |

dataset, and a test set that comprises the remaining 40% of the tweets. The training set includes a total of 1,086 tweets, of which 1,000 contain the authors' biographies, providing valuable context for analysis.

It is important to note that the distribution of class labels within this dataset is unbalanced. Out of the total, 879 tweets have been classified as containing reclaimed slurs, while only 207 tweets are labeled as not containing reclaimed language (Table 1). This significant disparity highlights the challenges posed by such an unbalanced dataset, emphasizing the need for careful consideration in analysis and modeling approaches.

# 5. Methods

## 5.1. EDA and Pre-processing

An exploratory data analysis (EDA) was first conducted on the training set to identify patterns and inform preprocessing decisions. Since usernames and URLs were already anonymized, they were removed along with redundant white spaces. A frequency analysis of hashtags and emojis was performed to assess their potential contribution to classification. Differences in relative frequencies between classes suggested that both emojis and hashtags carry informative signals for distinguishing reclamatory from non-reclamatory uses. Consequently, these elements were retained in the text.

Further frequency analysis was conducted on words in general and on characteristic terms to evaluate their association with each class. For this purpose, the HurtLex lexicon (Bassignana et al. 2018) was employed. HurtLex is a multilingual lexicon of offensive and derogatory expressions derived from Tullio De Mauro's *Le parole per ferire* and later expanded into a machine-readable resource for hate-speech research. The Italian seed lexicon groups over a thousand hate-related terms into 17 semantic categories, covering explicit slurs as well as stereotyped or contextually offensive descriptors (e.g., categories related to ethnicity, disability, sexuality, swear words, and crime). HurtLex was further enriched with part-of-speech information via MultiWordNet and with definitions and multilingual mappings via BabelNet, yielding a cross-lingual lexicon available in more than 50 languages. In this work, only the Italian terms belonging to the homophobic category we're selected and regular expressions were used to generate feminine, masculine, and plural variants. Results indicate that most slurs appearing in reclamation tweets (e.g., *frocio*, *gay*) also occur in non-reclamation contexts. However, there are two notable exceptions, which appear most in reclamation tweets. The first one is *frocia*, which, as noted by Nossem (2019) [17] and Cuccarini et al. (2024), represents a localized adaptation of the English term queer. Such localized forms are often created through reappropriation and semantic redefinition, making *frocia* highly indicative of reclamation and rarely used in a non-reclaimed sense. The second one is the orthographic variation *forci*, which is a community-specific marker of pride for the LGBTQ+ community.

Following the exploratory data analysis, the original training set was partitioned into training and validation subsets in an 85:15 proportion. Then, tokenization was performed using the tokenizer associated with the selected ELECTRA model, implemented via Hugging Face's `AutoTokenizer` class[1] to ensure compatibility with the pre-trained architecture. The column containing authors' biographies was preserved only for subtask B1.

[1] https://huggingface.co/docs/transformers/v4.14.1/en/model_doc/auto#transformers.AutoTokenizer

**Table 2**
Hyperparameter search space

| Subtask | Learning rate | Batch size | Weight Decay | Warmup steps | Accumulation steps |
|---|---|---|---|---|---|
| A1 | [1e-5, 5e-5] | $\{16, 32\}$ | $[0.01, 0.1]$ | $[0, 500]$ | None |
| B1 | [1e-5, 5e-5] | $\{8\}$ | $[0.01, 0.1]$ | $[0, 500]$ | $\{1, 2, 4\}$ |

### 5.1.1. Tweets and bios concatenation

To utilize both tweets and user bios as input for fine-tuning in subtask B1, the two fields are concatenated during tokenization. The tweet is treated as the first sequence, and the bio is treated as the second sequence, with the tokenizer encoding them in the format: `[CLS] tweet tokens [SEP] bio tokens [SEP]`. This structure allows the model to process both sequences simultaneously, leveraging the contextual information from both the tweet and the bio for classification tasks.

To handle the model's maximum input length, the tokenizer is configured to truncate only the bio if the combined length exceeds the limit, ensuring the tweet remains intact. This approach enables the model to learn the relationship between the two inputs and use them effectively for prediction.

## 5.2. Architecture and hardware

The system is built on an encoder-only transformer architecture, specifically an Italian variant of ELECTRA[2]. ELECTRA introduces a novel pre-training objective called Replaced Token Detection (RTD), which differs from BERT's Masked Language Modeling by training a discriminator to identify whether each token in the input is original or replaced by a generator (Clark et al. 2020). This approach provides training signals for every token rather than only the masked subset, making ELECTRA significantly more sample-efficient and enabling it to achieve better performance with fewer parameters compared to BERT-based models of similar size Empirical evaluations confirm that ELECTRA often outperforms BERT under equivalent computational budgets, particularly for smaller models, while maintaining strong contextual representations (Clark et al. 2020).

For this project, the Italian ELECTRA model was fine-tuned for binary sequence classification. The architecture consists of the pre-trained ELECTRA encoder and a classification head implemented via Hugging Face's `AutoModelForSequenceClassification`[3]. This head includes a dropout layer followed by a linear layer that outputs logits for two classes; softmax is applied only at inference to compute probabilities. Fine-tuning was performed using the Hugging Face Transformers library, leveraging the `Trainer` API[4] for streamlined training and evaluation. Hyperparameter optimization was integrated through Optuna[5], targeting learning rate, batch size, weight decay, and warmup steps, with macro F1-score as the optimization objective.

All the preprocessing and fine-tuning steps were performed with a local laptop Nvidia RTX 3060 GPU.

## 5.3. Hyperparameters tuning

For both subtasks, hyperparameter tuning was carried out using the Hugging Face Trainer API in combination with Optuna for automated optimization. The search spaces employed are reported in Table 2. Using identical search spaces for the two subtasks led to out-of-memory errors in subtask B1, primarily due to longer input sequences resulting from the concatenation of tweets and user bios. To mitigate this issue, the batch size search space for subtask B1 was reduced, and gradient accumulation steps were introduced as an additional parameter to be tuned. Furthermore, while 50 trials were executed for subtask A1, only 30 trials were performed for subtask B1. Each run was set to a maximum of 10

**Table 3**
Hyperparameters from best runs

| Subtask | Learning rate | Batch size | Weight Decay | Warmup steps | Accumulation steps |
|---|---|---|---|---|---|
| A1 | 2.35e-5 | 32 | 0.078 | 1 | None |
| B1 | 2.06e-5 | 8 | 0.014 | 291 | 4 |

**Table 4**
Test results

| Subtask | Macro F1 | Macro Precision | Macro Recall |
|---|---|---|---|
| A1 | 0.8706 | 0.9150 | 0.8395 |
| B1 | 0.8643 | 0.8938 | 0.8416 |

epochs, with early stopping applied to terminate unpromising runs.

After completing the search for each subtask, the hyperparameters from the best-performing trial (Table 3) were used to train a new model from scratch. Hardware constraints were again considered during this phase: a safe batch size was fixed, and gradient accumulation was employed to replicate the effective batch size of the optimal configuration without exceeding memory limits.

## 6. Results and discussion

### 6.1. Results

Table 4 reports the performance metrics for both subtasks. Overall, the model attains strong results in the textual classification setting (A1), with slightly lower performance when contextual information (author biographies) is incorporated (B1). However, these outcomes should be interpreted with caution: the hyperparameter search spaces and optimization strategies differed across subtasks due to memory constraints, which limits the validity of direct comparisons between A1 and B1.

In B1, the inclusion of biographies yields a marginal increase in recall for tweets that contain reclaimed language, but this comes at the cost of precision, producing a negligible change in macro F1. These findings suggest that simple concatenation of contextual data does not substantially improve performance under the current setup. Moreover, the dataset's limited size and pronounced class imbalance likely bias the classifier toward the majority class, further constraining generalization.

In short, while the architecture performs well in recognizing tweets that do not contain reclaimed language, reliably identifying tweets that contain reclamation language remains challenging, even with additional context. Future work should explore richer modeling of pragmatics and stance, alongside data augmentation and rebalancing strategies, to improve sensitivity to reclamatory language.

The leaderboard for Subtask A1 (Table 5)shows that the UniBO-FICLIT system ranks just below the official baseline, which is based on a BERT-style architecture fine-tuned with a weighted cross-entropy objective. This outcome aligns with earlier observations about the significant class imbalance in the dataset: because reclaimed tweets form only a small portion of the training data, optimizing with an unweighted loss tends to favor the majority class. Consequently, the baseline's superior performance appears to stem more from its training strategy than from inherent architectural differences. It is reasonable to assume that integrating class weights into the loss function of the system presented in this work would have improved the system's ability to identify reclaimed instances and potentially allowed it to reach the baseline's performance. In this sense, the A1 results reinforce one of the central methodological limitations highlighted in the thesis.

[2] https://huggingface.co/dbmdz/electra-base-italian-xxl-cased-discriminator

[3] https://huggingface.co/docs/transformers/v4.14.1/en/model_doc/auto#transformers.AutoModel

[4] https://huggingface.co/docs/transformers/main_classes/trainer

[5] https://huggingface.co/docs/transformers/main/en/hpo_train

**Table 5**
Participants results for Subtask A1

| Team | Run | F1-score |
|---|---|---|
| Ghavidel-Rajabi | 1 | 0.8981 |
| MilaNLP | 1 | 0.8959 |
| Ghavidel-Rajabi | 2 | 0.8909 |
| SaFe Tweets | 1 | 0.8895 |
| LlaNa | 1 | 0.8835 |
| GRUPPETTOZZO | 1 | 0.8834 |
| Challenger | 1 | 0.8816 |
| HateItOff | 1 | 0.8809 |
| GRUPPETTOZZO | 2 | 0.8735 |
| baseline | 1 | 0.8731 |
| UniBO-FICLIT | 1 | 0.8707 |
| NamDang | 2 | 0.8589 |
| HateItOff | 2 | 0.8584 |
| NetGuardAI | 1 | 0.8537 |
| The Hate Busters | 2 | 0.8503 |
| I2C | 2 | 0.8435 |
| NamDang | 1 | 0.8407 |
| Kenji-Endo | 1 | 0.8360 |
| The Hate Busters | 1 | 0.8345 |
| MilaNLP | 2 | 0.8306 |
| KIT-TIP-NLP | 2 | 0.8103 |
| Avahi | 1 | 0.7904 |
| KIT-TIP-NLP | 1 | 0.7659 |
| I2C | 1 | 0.6202 |

A different scenario emerges for Subtask B1 (Table 6). In this setting, the baseline achieves a very strong score, with only one system managing to surpass it. This suggests that the baseline constitutes a particularly robust point of comparison, and that the contextual configuration of B1 poses additional modeling challenges. The fact that nearly all participating systems fall short of the baseline indicates that well-regularized encoder models remain difficult to outperform when dealing with contextual inputs such as user biographies, especially under the constraints of this shared task.

## 6.2. Error analysis

Systematic misclassifications can be grouped into three recurrent phenomena: (i) orthographic variation in slurs, (ii) echoic attribution through quotation marks or reported speech, and (iii) the absence or ambiguity of contextual stance cues such as author metadata.

**Orthographic variation**. Several false negatives involve orthographic variations of slurs (e.g., *forcio* for *frocio*), which degrade the lexical signal on which the model relies. Subword splits become rare, embeddings are poorly estimated, and the distinction between reclaimed and non-reclaimed usage is obscured. These cases highlight the need for robust preprocessing strategies—such as character-level augmentation or fuzzy matching against a curated slur lexicon—and model components less sensitive to orthographic noise.

**Echoic attribution**. A second cluster includes tweets where slurs appear in quotation marks or are explicitly attributed to other speakers. In these cases, the author does not straightforwardly deploy the slur but uses it in an attributive way, representing an utterance or thought attributed to someone else (Bianchi 2014) [7]. More specifically, these are echoic uses, where the speaker signals an attitude toward the attributed content—often dissociative or mocking. While Bianchi argues that all reappropriation cases are echoic, quotation marks make this use more explicit. The model, however, seems to rely excessively on the presence of specific slur tokens and their direct use by the author, failing to capture

**Table 6**
Participants results for Subtask B1

| Team | Run | F1-score |
|---|---|---|
| LlaNa | 1 | 0.9021 |
| baseline | 1 | 0.8981 |
| GRUPPETTOZZO | 1 | 0.8979 |
| The Hate Busters | 2 | 0.8827 |
| MilaNLP | 1 | 0.8827 |
| GRUPPETTOZZO | 2 | 0.8681 |
| UniBO-FICLIT | 1 | 0.8644 |
| MilaNLP | 2 | 0.8641 |
| AIWizards | 1 | 0.8564 |
| AIWizards | 2 | 0.8549 |
| HateItOff | 2 | 0.8462 |
| The Hate Busters | 1 | 0.8324 |
| HateItOff | 1 | 0.8319 |
| SaFe Tweets | 1 | 0.8164 |
| Avahi | 1 | 0.8031 |
| DataSummit | 1 | 0.7734 |
| Kenji-Endo | 1 | 0.7489 |
| NetGuardAI | 1 | 0.7405 |

the echoic dimension of the reclamation phenomenon.

**Contextual stance cues**. Some misclassifications hinge on missing or ambiguous author metadata. For instance, tweets such as *"Avete paura di non potere più picchiare e insultare i 'froci'???"* are difficult to classify without knowing whether the author belongs to the target group. Although reclamation can also be performed by out-group speakers (Bianchi 2014; Cepollaro and López de Sa 2023) [8], the lack of clear contextual signals complicates interpretation. This difficulty also explains certain false positives, where the author's stance is ambiguous even for human annotators. For example, in *"Prima che le haters mi inizino a dire eH mA qUeStA è OmOfObIa, volevo comunicarvi che frocia lo sono pure io #circozzi"*, the slur is self-directed in a non-derogatory way, yet the initial framing suggests prior accusations of homophobia, making the overall intent unclear.

Additional false positives occur when the system over-relies on specific slurs (e.g., *frocio*) as indicators of reclamation, even in contexts lacking any reclaiming attitude, or when ambiguity arises from terms that can function as slurs in some contexts but not others. These cases underscore the complexity of reclamation and the limitations of purely lexical approaches, suggesting the need for models that integrate pragmatic and contextual reasoning.

The patterns observed collectively suggest that reclaimed language is not well captured by a binary decision boundary. As suggested by Bianchi (2014) and Cepollaro and López de Sa (2023) reclamation depends on many aspects (speaker identity, audience uptake, speaker attitude), many of which are gradient, context-relative, and multi-dimensional rather than strictly present/absent. Felicity and accomplishment conditions vary across different settings (Cepollaro and López de Sa 2023; Brontsema 2004) [4]. This complexity argues for modeling approaches beyond binary classification. Future datasets should reflect these dimensions to better align annotations with the phenomenon's structure.

## 7. Conclusion

This work presented an encoder-based approach to the MultiPRIDE task at EVALITA 2026, tackling the detection of reclaimed language in tweets containing LGBTQ+–related slurs. Fine-tuning ELECTRA yielded competitive performance, but persistent challenges remain—most notably under severe class imbalance and limited training data.

While the architecture performs well in recognizing tweets that do not contain reclaimed language, reliably identifying tweets that contain reclamation language is difficult, even when author biographies are included. The marginal changes observed in the contextual setting suggest that simple concatenation is insufficient to capture reclamation. Moreover, differences in hyperparameter search spaces across subtasks caution against direct comparisons.

The error analysis highlights three main sources of misclassification: orthographic variation, slurs' attribution, and ambiguous stance cues. Some of these findings suggest that reclamation is too complex to be faithfully modeled as a binary task. The phenomenon is inherently multi-dimensional and context-dependent, often requiring distinctions among stance, target, intent, felicity and accomplishment.

Overall, progress will likely depend on broadening the modeling perspective beyond binary outcomes, while concurrently strengthening the empirical foundation: more resilient preprocessing to handle orthographic variation and systematic exploration of data augmentation and rebalancing to address limited and skewed training distributions.

## Acknowledgments

The author thanks Professor Fabio Tamburini for his helpful advice during the development of this project.

## Declaration on Generative AI

During the preparation of this work, the author used GPT-5.2 in order to: Grammar and spelling check. After using the tool, the author reviewed and edited the content as needed and takes 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] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Comput. Surv. 51 (2018). URL: https://doi.org/10.1145/3232676. doi:10.1145/3232676.

[3] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523. URL: https://doi.org/10.1007/s10579-020-09502-8. doi:10.1007/s10579-020-09502-8.

[4] R. Brontsema, A queer revolution: Reconceptualizing the debate over linguistic reclamation, Colorado Research in Linguistics 17 (2014) 1–17. URL: https://www.sciencedirect.com/science/article/pii/S0378216614000526. doi:10.1016/j.pragma.2014.02.009.

[5] C. Hom, The semantics of racial epithets, Journal of Philosophy 105 (2008) 416–440. doi:10.5840/jphil2008105834.

[6] L. Anderson, E. Lepore, Slurring words, Noûs 47 (2013) 25–48. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0068.2010.00820.x. doi:10.1111/j.1468-0068.2010.00820.x.

[7] C. Bianchi, Slurs and appropriation: An echoic account, Journal of Pragmatics 66 (2014) 35–44. URL: https://www.sciencedirect.com/science/article/pii/S0378216614000526. doi:10.1016/j.pragma.2014.02.009.

[8] B. Cepollaro, D. López de Sa, The successes of reclamation, Synthese 202 (2023). URL: https://doi.org/10.1007/s11229-023-04393-y. doi:10.1007/s11229-023-04393-y.

[9] E. Zsisku, A. Zubiaga, H. Dubossarsky, Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination, in: Proceedings of the 16th ACM Web Science

Conference, WEBSCI '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 241–249. URL: https://doi.org/10.1145/3614419.3644025. doi:10.1145/3614419.3644025.

[10] M. Cuccarini, L. Draetta, C. Ferrando, L. James, V. Patti, Reclaim project: Exploring italian slurs reappropriation with large language models, in: F. Dell'Orletta, A. Lenci, M. Simonetta, S. Rachele (Eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), volume 3878 of *CEUR Workshop Proceedings*, CEUR-WS.org, Aachen, 2024, pp. 1–8. URL: https://ceur-ws.org/Vol-3878/39_main_long.pdf.

[11] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, Hodi at evalita 2023: Overview of the first shared task on homotransphobia detection in italian, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, Aachen, 2023. URL: https://ceur-ws.org/Vol-3473/paper26.pdf.

[12] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: E. Cabrio, A. Mazzei, F. Tamburini (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), volume 2253 of *CEUR Workshop Proceedings*, CEUR-WS.org, Aachen, 2018. URL: https://ceur-ws.org/Vol-2253/paper49.pdf.

[13] E. Lavergne, S. Rajkumar, G. Kovács, K. Murphy, Thenorth @ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, Aachen, 2020. URL: https://ceur-ws.org/Vol-2765/paper135.pdf.

[14] M. Sanguinetti, G. Comandini, E. di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, Aachen, 2020. URL: https://ceur-ws.org/Vol-2765/paper135.pdf.

[15] D. Locatelli, L. Locatelli, Lcts at hodi: Homotransphobic speech detection on italian tweets, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, Aachen, 2023. URL: https://ceur-ws.org/Vol-3473/paper30.pdf.

[16] V. Basile, M. Lai, M. Sanguinetti, Long-term social media data collection at the university of turin, in: E. Cabrio, A. Mazzei, F. Tamburini (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), volume 2253 of *CEUR Workshop Proceedings*, CEUR-WS.org, Aachen, 2018. URL: https://ceur-ws.org/Vol-2253/paper48.pdf.

[17] E. Nossem, Queer, frocia, femminielle, ricchione et al.- localizing 'queer' in the italian context, GSI: Gender, Sexuality, Italy 6 (2019) 1–27. URL: http://www.gendersexualityitaly.com/?p=2865.

## A. Online Resources

The source code of the project is available on github:

- https://github.com/simocasaz/evalita2026-multipride-UniBO-FICLIT,