MilaNLP at MultiPRIDE: Evaluating Lexical, Transformer, and Retrieval-Augmented Models for Reclaimed Language

Ivana Crescenzi[1], Arianna Muti[1] and Debora Nozza[1]

[1]*Bocconi University, Milan, Italy*

Abstract

This paper describes our participation in the **MultiPRIDE** shared task at EVALITA 2026, which focuses on the detection of reclaimed language within the LGBTQ+ community. The task aims to distinguish whether a potentially offensive word is used with a *reclaimed* intent or with a *not reclaimed* one. Our system combines TF–IDF features with logistic regression, achieving strong performance through the union of word and character n-grams that capture both explicit terminology and morphological variation. We also experiment with transformer-based models (AlBERTo), instruction-tuned language models (Qwen), and retrieval-augmented approaches to investigate whether external knowledge can enhance classification. Our results show that sparse lexical features combined with linear classifiers achieve very good performance (macro F1 = 0.8959 on the official test set). On the official leaderboard, our **constrained** TF–IDF submissions ranked **2nd** in **Task A** (text-only) and **5th** in **Task B** (text+bio), while our unconstrained retrieval-augmented runs ranked 20th in Task A and 8th in Task B. This highlights that reclaimed language in Italian social media relies heavily on explicit lexical signals.

Keywords

EVALITA, Italian NLP, hate speech, reclaimed language

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

1. Introduction

Online social media has dramatically expanded opportunities for communication, but it has also fostered the diffusion of hate speech. In particular, speech targeting sexual orientation, gender identity and gender expression remains a serious issue for LGBTQ+ communities, affecting well-being and inclusion. Despite progress in automatic hate speech detection, many challenges persist, especially the ones that concern more subtle usage of not reclaimed terms.

A particularly complex phenomenon is *reclaimed language* which is when members of marginalized groups reappropriate slurs and offensive terms, transforming them into expressions of identity, solidarity, or resistance. This linguistic practice, also known as semantic reappropriation or requalification [1], involves taking words that have historically been used as weapons of oppression and reclaiming them as tools of empowerment. The same word, such as *frocio*, *ricchione*, or *checca* in Italian, can be deeply offensive when used by outsiders to demean and marginalize, yet serve as a powerful expression of pride and community belonging when used by LGBTQ+ individuals in self-reference or in-group communication.

This duality poses a fundamental challenge for automated content moderation and hate speech detection systems. Standard classifiers that treat offensive terms as uniformly harmful risk silencing the very communities they aim to protect, censoring expressions of identity and solidarity while failing to address genuine hate. Conversely, systems that ignore hateful uses may allow harmful content to proliferate under the guise of community language. The distinction is not merely technical but carries significant social and ethical implications for freedom of expression, community autonomy, and online safety.

We address the task of classifying whether a textual instance uses a potentially offensive term with a *reclaimed* intent or a *non-reclaimed/derogatory* intent. Detecting reclamation requires models that can

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

ivana.crescenzi@unibocconi.it (I. Crescenzi); arianna.muti@unibocconi.it (A. Muti); debora.nozza@unibocconi.it (D. Nozza)

capture not just lexical patterns but also pragmatic context, authorial intent, and the subtle markers of in-group versus out-group language use.

The MultiPRIDE shared task frames this challenge as a binary text classification problem with two subtasks: **Task A** processes only the message text x, while **Task B** additionally incorporates the author's user biography b when available. For each instance, systems must predict whether the text employs potentially offensive terminology with a reclaimed intent ($y{=}1$) or with a not reclaimed intent ($y{=}0$). Participants can adopt either a *constrained setting* (using only the provided training data and lexical resources) or an *unconstrained setting* (allowing external annotated datasets). This structure enables direct comparison between text-only and context-aware approaches, and between resource-limited and resource-rich scenarios.

We proceed in five steps: (i) we perform a detailed lexical and length-based analysis of the dataset, revealing systematic differences between reclaimed and non-reclaimed usage patterns, (ii) we implement a baseline that combines TF–IDF features with a linear classifier, exploring word-level and character-level representations, (iii) we consider other state-of-the-art models including transformer-based architectures and instruction-tuned language models, (iv) we explore retrieval-augmented approaches that leverage external knowledge from related hate speech corpora, and (v) we discuss error patterns and open issues specific to reclaimed language detection, providing insights into the linguistic phenomena that challenge automated classification.

Our main takeaway is that simple, interpretable models based on TF–IDF representations can achieve high performance (macro F1 = 0.896) on the test set of this task. This result is particularly noteworthy given the complexity of the phenomenon and suggests that reclaimed language in Italian social media relies heavily on explicit lexical markers that sparse feature representations can capture effectively.

2. Background

Automatic detection of hate speech has been widely studied. Early and recent work explored multi-faceted textual representations (e.g. embeddings, sentiment, topicality) for social media classification [2]. Transformer-based models and domain-specific embeddings further improve performance across different benchmarks depending on the different tasks [3]. Beyond content-only methods, user-centric and network-aware approaches analyze interaction patterns and graph structure to characterize hateful users [4]. More recently, studies have specifically targeted homophobic and LGBTQ+-oriented hate, underscoring the importance of contextual modeling for nuanced decisions [5].

The phenomenon of *abusive versus non-abusive swearing* has received particular attention in recent years. Pamungkas et al. (2020) [6] highlighted the importance of distinguishing abusive from non-abusive uses of swear words in social media, demonstrating that the same lexical form can serve different communicative functions depending on context. They developed the SWAD (Swear Words Abusiveness Dataset) corpus for English, where abusive swearing is manually annotated at the word level, and showed that lexical, syntactic, and affective features are informative for automatic prediction of swearing functions. Their work emphasized that swearing is multifaceted and often serves positive social functions, including strengthening social bonds and improving conversation harmony, and explicitly noted the importance of recognizing reclaimed uses of slurs by marginalized communities. This line of research was further developed in their 2023 study [7], which provided a more comprehensive investigation of the role of swear words in abusive language detection tasks.

Most recently, and directly relevant to our work, Draetta et al. (2024) [8] presented the first study specifically addressing slur reappropriation detection in Italian. Using Large Language Models with zero-shot prompting approaches on a filtered subset of the HODI dataset [9], they explored whether instruction-tuned models could recognize reappropriative uses without supervised training. Their work is particularly important for highlighting the linguistic and cultural specificity of reclaimed language in Italian, where terms like *frocio*, *frocia*, *ricchione*, and *checca* undergo complex semantic shifts that differ from English reclamation patterns. They developed detailed annotation guidelines distinguishing friendly contexts, political reappropriation, and artistic uses, and found that even with carefully designed

prompts, the best-performing LLM configuration achieved only macro F1 = 0.58. Their analysis revealed moderate inter-annotator disagreement (Fleiss' Kappa = 0.57), underscoring the inherent difficulty of the task even for expert human annotators, and identified specific challenges such as the ambiguous status of *frocia* (which emerged from an already-reappropriative context) and the difficulty of determining speaker community membership from text alone.

However, the specific phenomenon of *reclaimed language* as a distinct detection task remains underexplored in mainstream datasets and systems; distinguishing self-referential reappropriation from not reclaimed usage is essential for fair and accurate moderation and analysis. The MultiPRIDE task is designed precisely to address this gap [10, 11], providing datasets in multiple languages and promoting the development of systems that can navigate this subtle distinction.

3. Data

The dataset provided consists of 1,086 Italian tweets, each associated with a binary label indicating whether the text contains reclaimed or non-reclaimed language (1 and 0, respectively). For each instance, we have access to the tweet text itself and, in approximately 92% of cases, also the author's short biography.

Split	#Docs	Not Reclaimed	Reclaimed
Full	1,086	879	207
Train+Dev	868	703	165
Validation	218	176	42

Table 1
General statistics and stratified data split.

The dataset is imbalanced (approx. 1:4), with relatively short and homogeneous texts. Samples that use reclaimed language are shorter with more explicit slurs while not-reclaimed instances (which include both derogatory and neutral/argumentative discussions) tend to be longer and more discursive(see Appendix A, Table A).

We divide the dataset into a stratified `train_dev` split (80%) and a `validation` set (20%). Stratification preserved the 1:4 class ratio between reclaimed and not reclaimed examples. We adopt a five-fold cross-validation protocol with out-of-fold (OOF) aggregation to assess robustness and minimize variance.

4. Task Description

The MultiPRIDE task is designed as a binary text classification problem: given an Italian social media message x and, when available, a user biography b, the goal is to predict whether the text employs potentially offensive terminology with a *reclaimed* intent (y=1) or an *not reclaimed* intent (y=0). This differs from generic hate speech detection because it requires recognizing self-identification and irony. The task offers two tasks: **Task A** processes only the message text x, while **Task B** combines text x with biography b. For both tasks, participants can adopt either a *constrained* setting (using only the provided training data) or an *unconstrained* setting (allowing external datasets). We normalize all texts through a minimal pipeline: (1) lowercasing; (2) replacement of mentions and URLs with the placeholders <USER> and <URL>; and (3) whitespace normalization. We deliberately preserve punctuation and emojis since they often signal stance, irony, or emotional emphasis. For Task B (contextual setting), we concatenated biographies to tweets as `text [SEP] bio`, allowing a single model to process both sources without a separate encoder.

5. Methodology

As a first baseline, we train various linear models based on **TF–IDF** features combined with **Logistic Regression**. These approaches offer high interpretability and robustness. We test different lexical granularities, capturing both word-level cues and subword patterns. All models employ class weighting to mitigate the dataset imbalance, and their behavior can be directly interpreted in terms of the most predictive lexical features. We report detailed parameter configurations in the Appendix B.

To benchmark more expressive encoder-only architectures, we fine-tuned the Italian BERT variant **AlBERTo** (`dbmdz/bert-base-italian-uncased`) under two input configurations: (i) tweet text only and (ii) tweet text concatenated with the author's biography (using `[SEP]` as separator).

We fine-tune AlBERTo adopting stratified 5-fold cross-validation and a class-weighted objective to mitigate imbalance. Full training hyperparameters are reported in Appendix B.

Text and biography were concatenated only when a non-empty biography was available:

$$\text{input} = \text{text} \parallel \texttt{[SEP]} \parallel \text{bio}.$$

We further evaluate two instruction-tuned generative models from the **Qwen Instruct** series, with 1.5B and 7B parameters, respectively. Both models were used *exclusively via prompting*, without any fine-tuning, with the goal of assessing whether general-purpose LLMs can identify reclamation phenomena through instruction following alone.

Specifically, the **Qwen Instruct 1.5B** model was employed for the *text-only* task, as this setting primarily requires linguistic understanding and can be effectively addressed by a smaller-capacity model. In contrast, the **Qwen Instruct 7B** model was adopted for the *text + bio* task, where the increased informational complexity arising from the integration of biographies demands stronger representation and reasoning capabilities, which are better supported by a larger model.

We evaluate Qwen Instruct models using a deterministic decoding setup to force binary outputs. The prompts are specified in the Appendix C and, since the tweets are in Italian, the prompts are also in Italian. Full inference settings are reported in Appendix B.

5.1. External Knowledge

To leverage external knowledge without supervised fine-tuning, we develop a retrieval-augmented pipeline using the **HODI** Subtask A corpus [9] as external knowledge base. Our approach proceeds in three stages: (i) a kNN label propagation baseline to assess direct label propagation trough retrieval of nearest neighbours, (ii) a text retrieval mechanism that extracts contextual evidence rather than labels, and (iii) a classification pipeline that combines retrieved context with the original features. All texts undergo consistent preprocessing: lowercasing, entity masking (@user → <USER>, URL → <URL>), and whitespace normalization.

kNN Label Propagation Baseline. We first evaluate direct label propagation via k-Nearest Neighbors (kNN) with $k = 3$. Similarity is computed using TF-IDF cosine distance on normalized HODI texts. Each test instance receives the majority label of its top-3 neighbors, weighted by cosine similarity.

This baseline yields macro F1 = 0.55 on validation but reveals critical limitations:

- Reclaimed recall = 0.26: amplifies HODI's class imbalance.
- Many reclaimed instances are misclassified as not reclaimed because they share lexical cues (especially slurs) with hate speech and this issue is not properly addressed by the HODI task which targets a different objective.

Text Retrieval Mechanism. We address this limitation that emerges in the kNN approach by retrieving *textual evidence* rather than labels through majority vote. For each input q, we again retrieve top-$k = 3$ HODI instances via a dual TF-IDF encoder capturing lexical and morphological patterns. Retriever design choices and parameterization are reported in Appendix B.

We pre-compute TF–IDF for all HODI instances using word 1–2 grams and character 2–5 grams. Given a test query, we build its TF–IDF vector with the same features and retrieve the top-k nearest neighbors by cosine similarity. In addition to the retrieved textual context, we incorporate similarity-based signals summarizing the composition and relevance of the retrieved neighborhood. Implementation details are reported in Appendix B. We move from direct neighbor label transfer to a supervised classifier trained on **MultiPRIDE**. For each MultiPRIDE instance, we retrieve the top-k most similar HODI texts in the same TF–IDF space. Rather than copying HODI labels, we convert the retrieved neighborhood into additional features and train a class-weighted Logistic Regression on MultiPRIDE `train_dev` labels. Concretely, the final vector concatenates: (i) TF–IDF features of the original input (tweet, or tweet+bio), (ii) TF–IDF features of a retrieved *context* obtained by concatenating (and truncating) the top-k HODI texts, and (iii) retrieval statistics such as mean/max cosine similarity and a neighborhood-composition signal. Full feature definitions and optimization details are reported in Appendix B.

6. Results

Table 2 summarizes cross-validation performance on `train_dev` (80%), while Table 3 shows results on our internal validation set (20%). Official test set results provided by the task organizers are reported in Section 6.2. For evaluation, we use **Macro F1**, Precision, and Recall, complemented by confusion matrices.

6.1. Held-out Validation Set

Linear models built on TF–IDF representations provide strong and consistent baselines. Among the variants tested, combining word- and character-level features yields the highest and most stable results. Using both tweet text and biography slightly improves the out-of-fold (OOF) performance, reaching a Macro F1 of 0.90 on average (Table 2). Notably, the purely lexical models perform remarkably well despite their simplicity.

Model	Macro F1 (mean±std)	OOF Confusion Matrix
TEXT: Word+Char_WB	0.8853 ± 0.0266	$\begin{bmatrix} 685 & 18 \\ 40 & 125 \end{bmatrix}$
TEXT+BIO (Word+Char_WB, Char)	0.9037 ± 0.0205	$\begin{bmatrix} 693 & 10 \\ 38 & 127 \end{bmatrix}$
TEXT+BIO (word + Char_WB, Char_WB)	0.9037 ± 0.0205	$\begin{bmatrix} 693 & 10 \\ 38 & 127 \end{bmatrix}$
TEXT+BIO (Char, Char)	0.8769 ± 0.0400	$\begin{bmatrix} 690 & 13 \\ 47 & 118 \end{bmatrix}$

Table 2
Cross-validation (OOF, 5-fold) results on `train_dev`.

6.1.1. Validation Performance

Results on the validation set confirm the strong performance of lexical representations (Table 3). The text-only TF–IDF union achieves the best Macro F1 (0.9565), outperforming the text+bio variant (0.9023), which suggests that biography concatenation may introduce noise. AlBERTo reaches competitive but lower scores (0.8751 text-only, 0.8502 text+bio), indicating that the limited dataset size may constrain fine-tuning benefits. Qwen Instruct models perform significantly worse (0.1740 for 1.5B, 0.4304 for 7B), confirming that zero-shot prompting alone cannot reliably capture reclamation phenomena. Finally, retrieval variants improve over the kNN baseline (0.5461) but remain below internal TF–IDF models (0.8949 text-only, 0.8725 text+bio), suggesting partial but incomplete alignment between HODI and MultiPRIDE distributions.

Model	Macro Precision	Macro Recall	Macro F1	Confusion Matrix
TF–IDF Word(1-2)+CharWB(2-5) **TEXT only**	0.9488	0.9648	**0.9565**	$\begin{bmatrix} 172 & 4 \\ 2 & 40 \end{bmatrix}$
TF–IDF Word(1-2)+CharWB(2-5) + **BIO(Char)**	0.9548	0.8662	**0.9023**	$\begin{bmatrix} 175 & 1 \\ 11 & 31 \end{bmatrix}$
AlBERTo (TEXT)	0.9003	0.8548	0.8751	$\begin{bmatrix} 171 & 5 \\ 11 & 31 \end{bmatrix}$
AlBERTo (TEXT+BIO)	0.8788	0.8282	0.8502	$\begin{bmatrix} 170 & 6 \\ 13 & 29 \end{bmatrix}$
Qwen Instruct 1.5B (TEXT)	0.5972	0.5057	0.1740	$\begin{bmatrix} 2 & 174 \\ 0 & 42 \end{bmatrix}$
Qwen Instruct 7B (TEXT+BIO)	0.5273	0.5413	0.4304	$\begin{bmatrix} 69 & 107 \\ 13 & 29 \end{bmatrix}$
kNN Retrieval Only	0.5466	0.5457	0.5461	$\begin{bmatrix} 146 & 30 \\ 31 & 11 \end{bmatrix}$
TEXT ONLY + Retrieval	0.9025	0.8877	**0.8949**	$\begin{bmatrix} 170 & 6 \\ 8 & 34 \end{bmatrix}$
TEXT+BIO + Retrieval	0.9086	0.8458	**0.8725**	$\begin{bmatrix} 172 & 4 \\ 12 & 30 \end{bmatrix}$

Table 3
Final results on the `validation` set (Macro averages).

6.2. Official Test Set Results

On the official leaderboard, our **constrained** TF–IDF submissions ranked **2nd** in **Task A** (text-only) and **5th** in **Task B** (text+bio). In the **unconstrained** setting, our retrieval-augmented runs ranked **20th** in **Task A** and **8th** in **Task B**.

The evaluation reports released by the organisers include precision, recall, and F1 scores for each class. For both tasks, the constrained setting uses TF–IDF features only, while the unconstrained setting extends this configuration with a retrieval-augmented approach using the external HODI corpus.

6.2.1. Quantitative overview.

For Task A (text-only), the constrained system achieves a macro F1 of 0.8959, with very strong performance on class 0 (F1 = 0.9609) and significantly lower performance on class 1 (F1 = 0.8309). The unconstrained run shows a consistent drop in the reclaimed class (F1 = 0.7239) and in macro F1 (0.8306), suggesting that adding HODI data introduces domain mismatch effects that predominantly harm the reclaimed label. The same asymmetry holds in Task B (text+bio), although overall scores are higher. In particular, the constrained run obtains a macro F1 of 0.8826 with an excellent F1 of 0.9590 on class 0 but only 0.8063 on class 1. The unconstrained counterpart performs worse on the reclaimed class (F1 = 0.7769). Across both tasks, the model systematically achieves very high precision and recall for not reclaimed instances, typically above 0.94 for both metrics, and substantially lower recall for reclaimed instances. This indicates a pronounced bias toward the majority class, resulting in a high number of false negatives and makes reclamation errors the most critical category.

6.2.2. Impact of constrained vs. unconstrained training.

In both tasks, the constrained systems outperform the unconstrained ones, especially on the reclaimed class. While additional training data from HODI increases lexical variety, it introduces noise: many HODI examples contain explicit slurs used non-reclaimedly, shifting the decision boundary toward class 0 and decreasing the model's sensitivity to reclaimed uses in the MultiPRIDE distribution. This is particularly evident in the recall for class 1, which drops from 0.8129 (Task A constrained) to 0.6978 (Task A unconstrained), and from 0.7338 (Task B constrained) to 0.7266 (Task B unconstrained).

Thus, constrained training appears more aligned with the annotation style and pragmatic subtleties of MultiPRIDE.

7. Analysis of Retrieval Results

As reported in Table 3, the full retrieval pipeline substantially outperforms the kNN baseline, with improved recall on the reclaimed class due to contextual disambiguation rather than biased label transfer.

To better understand how the retrieval behaves on specific instances, we examine three representative cases where the retrieval either successfully disambiguates complex pragmatic contexts or fails due to domain mismatch between the MultiPRIDE distribution and the external HODI corpus. These examples illustrate both the potential advantages of textual contextualization over simple label voting and the limitations imposed by distributional misalignment.

Semantic disambiguation : "*@USER @USER @USER si sa che voi LGBT (froci non si può dire) ragionate col culo...*"
(English: "@USER @USER @USER everyone knows that you LGBT (you're not allowed to say 'faggots') think with your ass...")

This is a hateful tweet that uses the slur "froci" while simultaneously employing irony through the parenthetical remark "froci non si può dire" ("one cannot say faggots"), which mockingly references political correctness norms. The speaker uses the offensive term while pretending to acknowledge that it should not be used, adding a layer of sarcastic aggression. The text is correctly labeled as non-reclaimed because it deploys the slur to demean LGBTQ+ individuals, not as an act of self-identification or in-group solidarity. However, the meta-linguistic commentary ("non si può dire") and ironic framing make this instance pragmatically complex. The pure kNN retrieval baseline identifies two out of three nearest neighbors from HODI as hateful content and one as reclaimed content, leading it to correctly predict the hateful label by majority voting. However, this decision relies on frequency patterns and does not model the ironic nuance. In contrast, the retrieval classifier, beyond considering the 67% hateful neighbors, also uses the semantic content of the 33% reclaimed neighbor to contextualize the irony and confirm the hateful label more robustly.

Corpus misalignment : "*Le persone trans non sono ideologia sono esseri umani...*"
(English: "Trans people are not an ideology; they are human beings...")

This text represents a neutral rights-based discourse defending transgender people, but it does not contain reclaimed slurs, therefore it is correctly labeled as non-reclaimed. However, when the retriever searches for similar texts in the external HODI corpus, it encounters a fundamental issue: HODI is a homotransphobia detection dataset that predominantly contains hateful messages attacking LGBTQ+ individuals. The retriever matches on keywords like "trans" and "persone" (people), but the retrieved HODI examples tend to express prejudice rather than support. This introduces misleading signals: the original MultiPRIDE text is pro-LGBTQ+ and supportive, while the retrieved HODI context is anti-LGBTQ+ and hateful. The root cause is distribution mismatch: HODI's bias toward hateful content means it cannot adequately represent MultiPRIDE's non-reclaimed category, which includes both hate and neutral rights-based discussions. This mismatch explains why the retrieval provides limited improvement and can sometimes degrade performance.

Stylistic mismatch : "*L'attivismo trans e #queer nasconde #sessuofobia...*"
(English: "Trans and #queer activism hides #sexophobia...")

The discourse originates from within the LGBTQ+ community itself and reflects internal debates about political strategies, ideological positions, or priorities. While the text discusses LGBTQ+ topics and may contain terminology associated with reclaimed language contexts, it is not using slurs in a self-referential or reappropriative manner. Instead, it expresses critique in an argumentative register that differs from external homophobic attacks. When the retriever searches for similar texts in the

HODI corpus, it often retrieves generic slur-based hate speech from hostile outsiders. The problem is that surface-level TF-IDF similarity focuses on lexical overlap (terms like “trans”, “queer”, “activism”) while missing the pragmatic distinction between intra-community debate and out-group hostility. As a result, the retrieved HODI contexts provide limited useful information and may even introduce noise. This domain mismatch is amplified by HODI’s skew toward external hate, whereas MultiPRIDE’s non-reclaimed class also includes neutral discussions, political debates, and intra-community critiques, explaining why the retrieval yields limited benefit despite its theoretical advantage over label-based kNN voting.

8. Discussion

Word and character TF–IDF (with word boundaries) capture complementary evidence: word n-grams encode explicit offensive terms and short expressions, while character n-grams capture morphological variants and orthographic creativity. Their union achieves the best trade-off between precision and recall and is remarkably stable across folds. Contrary to the initial hypothesis that biography could help out, biographies can add self-identification cues but often introduce noise: they contain hashtags, quotes, and political statements unrelated to reclamation. The observed drop from text-only to text+bio on the test set demonstrates that a naive concatenation does not generalize. Selective conditioning, or profile encoders, may leverage this information more effectively.

AlBERTo delivers consistent results and demonstrates its robustness for Italian classification, but it remains slightly below the TF–IDF baseline. The dataset’s short and lexically explicit nature favors sparse representations. Fine-tuning on limited data also risks overfitting despite class-weighted loss. As a prompt-based alternative, the Qwen Instruct models underperform significantly (Macro-F1 0.17–0.43), producing unstable outputs and invalid labels having a precision of 0.59. This highlights the difficulty of using prompt-based inference alone for context-dependent linguistic phenomena such as reclamation.

The retrieval experiments confirm that naive retrieval from external sources can complement internal features but must be carefully aligned. The kNN vote (0.5461 Macro-F1) alone is insufficient, but once contextualized, recall on reclaimed uses improves. However, external content from HODI shows domain mismatch and weak similarity (cosine scores mostly < 0.2). This partial alignment explains why retrieval method increases recall but slightly decreases overall F1 relative to the text-only baseline.

Qualitative analysis. The following qualitative analysis refers to the **official test set** released by the task organizers and is based on the errors produced by our **constrained TF–IDF (text-only) submission**. A manual inspection of the 20 misclassified tweets provided by the task organizers reveals recurrent pragmatic configurations that explain the systematic recall drop on class 1. Five tweets contain reclaimed slurs embedded in aggressive, ironic or emotionally charged exchanges, where confrontational polarity overrides reclamation cues. This pattern is observed, for instance, in tweets such as *it_326* (“non è tanto per noi froci”; ”it’s not much for faggots”) and *it_913* (“urlo frocio a tutti”; ”I shout faggot to everybody”), where reclaimed terms occur within heated or sarcastic discourse. Five tweets reproduce homophobic slurs by explicitly quoting them in order to criticise, parody or contextualise hate speech. Examples include *it_418*, which reports “Semo tutti froci”(”we are all faggots”), and *it_112*, where the speaker refers to offensive remarks addressed to the LGBTQ+ community. The classifier fails to detect the reported or meta-discursive stance and focuses on the surface presence of the slur. Three tweets use reclaimed identity terms in low-marked self-descriptive contexts, where the term functions as bare self-identification rather than as an explicit positive evaluation. For example, in *it_524* (“frocia lo sono pure io”; ”I am also a faggot”) and *it_815* (“frocie, fluide, queer”; ”faggots, fluids, queer”), reclaimed labels are used descriptively, leading the model to predict class 0. Four tweets instantiate dialogic structures that require conversational context to interpret reclamation correctly. This includes direct replies or mentions such as *it_50* (“Avete paura di non potere più picchiare e insultare i ‘froci’?” ; ”Are you scared you can no longer insult faggots and beat them up”) and *it_986* (addressing @USER), where speaker stance and pragmatic intent depend on preceding turns. Without conversation-level modelling,

the system cannot capture reclamation emerging across turns. The combination of quantitative and qualitative evidence suggests that the model is highly competent in detecting prototypical offensive uses and that reclaimed uses are more heterogeneous and pragmatically marked, making them more difficult to capture.

9. Conclusions

In this work we described our participation in the MultiPRIDE task at EVALITA 2026, focusing on the detection of reclaimed language in Italian social media. Our experiments show that simple, interpretable models based on TF–IDF features combined with linear classifiers can outperform more complex neural architectures. These results highlight the strong role of explicit lexical cues in distinguishing reclaimed from non-reclaimed uses within this domain specifically relevant to the Italian language.

The analysis further indicates that additional contextual signals, such as user biographies or external retrieved examples, do not necessarily improve performance and may introduce noise when not carefully integrated. A limitation of our comparison is that the two tasks were evaluated using different model sizes, preventing a fully controlled assessment of the biographical context's contribution. The systematic error patterns observed across subtasks reveal the intrinsic difficulty of identifying reclaimed usage, especially in cases involving quotation, mixed-polarity contexts, or pragmatically nuanced in-group communication.

Declaration on Generative AI

Generative AI tools were used to assist with proofreading and structural suggestions during the writing of this paper. The authors take full responsibility for the final content.

References

[1] Bianchi, C. (2014). *Slurs and appropriation: An echoic account.* Journal of Pragmatics, 66:35–44. [page:1]

[2] Cao, R., Lee, K.-W. R., & Hoang, T.-A. (2021). *DeepHate: Hate Speech Detection via Multi-Faceted Text Representations.* arXiv:2103.11799.

[3] Saleh, H., Alhothali, A., & Moria, K. (2021). *Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model.* arXiv:2111.01515.

[4] Ribeiro, M. H., Calis, P. H., Santos, Y. A., Almeida, V. A. F., & Meira Jr., W. (2018). *Characterizing and Detecting Hateful Users on Twitter.* In Proc. ICWSM 2018, pp. 10–19.

[5] McGiff, J., & Nikolov, N. S. (2024). *Bridging the gap in online hate speech detection: A comparative analysis of BERT and traditional models for homophobic content identification on X/Twitter.* arXiv:2405.09221.

[6] Pamungkas, E. W., Basile, V., & Patti, V. (2020). *Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media.* In *Proc. LREC 2020*, pp. 6237–6246.

[7] Pamungkas, E. W., Basile, V., & Patti, V. (2023). *Investigating the role of swear words in abusive language detection tasks. Language Resources and Evaluation*, 57(1):155–188.

[8] Draetta, L., Ferrando, C., Cuccarini, M., James, L., & Patti, V. (2024). *ReCLAIM Project: Exploring Italian Slurs Reappropriation Detection with LLMs.* In *Proc. 5th CLiC-it 2024.*

[9] Nozza, D., Cignarella, A. T., Damo, G., Caselli, T., & Patti, V. (2023). *HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian.* In *EVALITA 2023: Technical Reports.* CEUR Workshop Proceedings, vol. 3473.

[10] 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 Lan-

guage Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy,2026.

[11] 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.

A. Exploratory Analysis

Text length. Across all samples, the average message length is **35 tokens** (SD 11.2) and **199 characters** (SD 61). Texts labeled as 0 are generally longer (mean **36.8** tokens; **216** characters) compared to those labeled as 1 (mean **22.5** tokens; **127** characters).

	Characters		Tokens	
Label	Mean	SD	Mean	SD
0	216.2	51.6	36.8	9.2
1	127.3	67.0	22.5	11.8

Table 4
Average text length per class (characters and tokens).

The token-length histogram shows a *mostly unimodal* distribution with a clear mode between **35–45** tokens, and a short left tail.

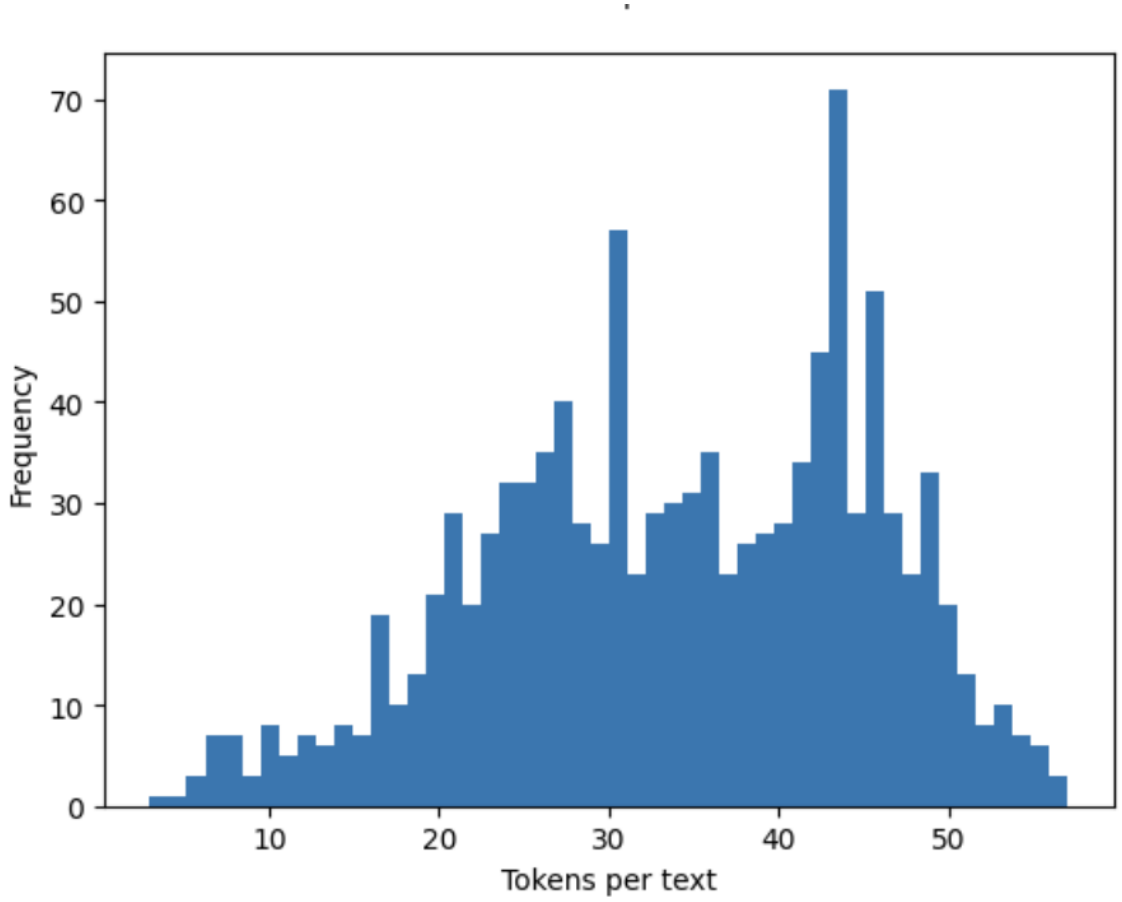

Figure 1: Distribution of token lengths across texts in the dataset.

Token frequency analysis (after stopword removal) reveals domain-coherent terms such as *trans*, *transfobia*, *donne*, *diritti*, *lgbt*, *gay*, together with placeholders like @USER and <URL>. Reclaimed samples (`label=1`) contain higher frequencies of commonly offensive terms (*frocio*, *froci*, *ricchione*), while offensive samples (`label=0`) tend to include more discursive or argumentative language (e.g., *diritti*, *transfobia*, *donne*).

Minor artifacts were detected (URLs, @USER mentions, placeholders), but no duplicated texts or IDs. A light normalization was applied: lowercasing, placeholder standardization (@USER, <URL>),

and whitespace cleaning. All experiments used a fixed random seed (42) for stratification and cross-validation.

B. Implementation Details and Hyperparameters.

B.1. TF–IDF and Linear Classifiers specifics.

Three complementary vectorizers were explored:

- **Word-level TF–IDF:** word n-grams (n=$\{1, 2\}$), sublinear term frequency, smoothed IDF, and `min_df` $= 2$, `max_df` $= 0.9$. This configuration captures both isolated slurs and short collocations.
- **Character-level TF–IDF:** two analyzers: (i) plain CHAR with n-grams of length 3–5, and (ii) CHAR_WB (word-boundary aware) with n-grams 2–5. These extract subword patterns robust to spelling variation, inflection, and creative obfuscation.
- **Feature Union:** concatenation of the Word(1–2) and Char_WB(2–5) spaces, followed by a class-weighted Logistic Regression.

B.2. AlBERTo Training Hyperparameters

The training hyperparameters defined in the official scripts were:

- **Batch size**: 16
- **Learning rate**: 2×10^{-5}
- **Epochs**: 3
- **Max sequence length**:
  - 160 for the *text-only* setting
  - 200 for the *text+bio* setting
- **Optimizer**: AdamW (HuggingFace default)
- **Seed**: 42
- **Evaluation strategy**: evaluation at every epoch
- **Loss**: dynamically weighted cross-entropy

B.3. Qwen Inference Hyperparameters

The main hyperparameters used in the inference scripts were:

- **Batch size**: 8
- **Max input length**:
  - 160 for *text-only*
  - 200 for *text+bio*
- **Max new tokens**: 2 (forcing the model to output only 0 or 1)
- **Decoding**: deterministic (`do_sample = false`, `temperature = 0.0`)
- **Few-shot**: random sampling of k labelled instances

B.4. Summary of Model Configurations

B.5. Logistic Regression Settings (Retrieval)

The classifier is trained with inverse-frequency class weighting, a regularization parameter $C = 1.0$, a maximum of 2000 iterations, and a linear optimization solver.

Model	Batch size	LR	Max length
AlBERTo (text)	16	2×10^{-5}	160
AlBERTo (text+bio)	16	2×10^{-5}	200
Qwen 1.5B (text)	8	—	160
Qwen 7B (text+bio)	8	—	200

Table 5
TF-IDF Retriever Configuration.

Component	Parameters
Word n-grams	(1–2), min_df=2, max_df=0.90, sublinear_tf
Char word-boundary	(2–5), min_df=2, sublinear_tf
Metric	Cosine similarity
Truncation	300 characters per retrieved text

B.6. Retriever Configuration

B.7. Similarity-based Features (Retrieval)

Beyond the retrieved textual context, we extract additional similarity-based signals to summarize the retrieved neighborhood. These include:

- the fraction of reclaimed instances among the top-k retrieved texts;
- the average cosine similarity between the query and retrieved instances;
- the maximum cosine similarity observed in the neighborhood.

The final representation stacks multiple views:

$$[\text{TF-IDF}(\text{text}),\ \text{TF-IDF}(\text{ctx\_text}),\ [\text{ctx\_pos\_frac}, \text{ctx\_sim\_mean}, \text{ctx\_sim\_max}]]$$

C. Prompts Used

Prompts for Qwen Instruct 1.5B (text-only).

> **Prompt:**
> Classifica il testo come 1 se contiene un uso riappropriato di termini LGBTQ+, oppure 0 se non riappropriato. Rispondi solo con 0 o 1.

Prompt for Qwen Instruct 7B (text+bio).

> **Prompt:**
> Sei un classificatore per il task MultiPRIDE (italiano). Ricevi un testo e, se presente, la bio del profilo (separata da [SEP]). Decidi se il termine è usato con riappropriazione (1) oppure no (0). Rispondi solo con 0 oppure 1.