MultiPRIDE at EVALITA 2026: Overview of the Multilingual Automatic Detection of Slur Reclamation in the LGBTQ+ Context Task

Chiara Ferrando[1,*,†], Lia Draetta[1,*,†], Marco Madeddu[1,*,†], Mae Sosto[2,*,†], Viviana Patti[1], Paolo Rosso[3,4], Cristina Bosco[1], Jacinto Mata[5] and Estrella Gualda[6]

[1]*University of Torino, Computer Science Department, Corso Svizzera 185, 10149 Torino, Italy*

[2]*Human Center Data Analytics Group, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands*

[3]*Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Camí de Vera, 46022 València, Spain*

[4]*Valencian Graduate School and Research Network of Artificial Intelligence, Camí de Vera, 46022 València, Spain*

[5]*Departamento de Tecnologías de la Información, Universidad de Huelva, Campus El Carmen Avda. de las Fuerzas Armadas. - 21007 Huelva, Spain*

[6]*Departamento de Sociología, Trabajo Social y Salud Publica, Universidad de Huelva, Campus El Carmen Avda. de las Fuerzas Armadas. - 21007 Huelva, Spain*

Abstract

MultiPRIDE is a shared task proposed at EVALITA 2026. The main focus of the task is the detection of slur reclamation, a linguistic phenomenon in which derogatory terms are reclaimed by members of a community. The challenge is divided in two main tasks: (A) binary detection of reclamation in social media posts and (B) contextual detection of reclamation considering users' biographies provided within the texts. As we propose a multilingual corpus, each task is further divided into language-specific subtasks (Italian, Spanish and, English). The task has been one of the most participated at EVALITA 2026 [1] with 19 teams submitting a total of 61 runs for Task A and 27 runs for Task B. In this paper, we present the dataset, the evaluation, the participating systems, and a discussion of the results.

All the code is available at https://github.com/multipride-evalita/MultiPRIDE-Task

Keywords

Hate Speech, Slur Reclamation, LGBTQ+, Social Media, Multilingual

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

1. Motivation

Hate Speech (HS) is intrinsically context-dependent and resists rigid classification, as abusive language can manifest in diverse forms, including insults, hostility, offensiveness, or disrespect [2, 3, 4].

A particularly underexplored dimension of HS concerns the **reclamation of slurs**, a linguistic practice whereby historically derogatory terms are reclaimed to express a sense of belonging and solidarity among in-group members [5, 6].

From a content moderation perspective, accurately identifying reclaimed uses of slurs is crucial to avoid further marginalizing communities that employ such terms as expressions of identity and belonging. Misclassification may restrict their freedom of expression [7, 8, 9] and lead to the paradoxical outcome of censoring forms of counter-speech [10].

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.

chiara.ferrando@unito.it (C. Ferrando); lia.draetta@unito.it (L. Draetta); marco.madeddu@unito.it (M. Madeddu); mae.sosto@cwi.nl (M. Sosto); viviana.patti@unito.it (V. Patti); prosso@dsic.upv.es (P. Rosso); cristina.bosco@unito.it (C. Bosco); mata@uhu.es (J. Mata); estrella@uhu.es (E. Gualda)

009-0000-9593-2510 (C. Ferrando); 0009-0004-6479-5882 (L. Draetta); 0009-0004-5620-0631 (M. Madeddu); 0009-0000-6252-2275 (M. Sosto); 0000-0001-5991-370X (V. Patti); 0000-0002-8922-1242 (P. Rosso); 0000-0002-8857-4484 (C. Bosco); 0000-0001-5329-9622 (J. Mata); 0000-0003-0220-2135 (E. Gualda)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To the best of our knowledge, slur reclamation remains largely underexplored in Natural Language Processing (NLP) [11], with only Draetta et al. [12] addressing the phenomenon beyond English. Against this background, the **MultiPRIDE task at EVALITA 2026** introduces a multilingual challenge aimed at distinguishing derogatory uses of slurs from their reclaimed uses in LGBTQ+ contexts. The proposed task seeks to advance research on this nuanced aspect of hate speech detection by fostering participation in the identification of reclaimed uses of slurs in a multilingual setting (English, Italian, and Spanish). To this end, we introduce the first multilingual corpus specifically dedicated to slur reclamation within the LGBTQ+ community, comprising social media texts in three languages (English, Italian, and Spanish) and drawn from multiple resources [13, 11, 14]. The dataset includes both social media posts and author biographies for the Italian and Spanish subsets. By incorporating biographical information, the task extends beyond text-level classification to investigate whether user-level information can support the detection of slur reclamation, a phenomenon that is inherently pragmatic and highly subjective. Finally, the shared task enables a systematic analysis of cross-linguistic and cultural variation in slur reclamation practices and facilitates the analysis of how models operating in different languages vary in their ability to address this challenge.

2. Definition of the Task

The task aims to encourage participants to investigate linguistic phenomena and challenges associated with reclaimed language within the LGBTQ+ community. In particular, participants are invited to analyze both the textual properties of the input data (such as argumentative strategies, slurs, denigratory expressions, self-referential labels, and figures of speech), and the contextual information that may be inferred from users' profiles (when available), including indicators of LGBTQ+ community membership or political orientation. We propose a **binary classification task** in which systems are required to determine whether a term related to the LGBTQ+ domain, as it appears in a sentence, is used with a reclamatory intent or not. The shared task is structured into two main tasks, each further divided into subtasks:

Task A - Slur Reclamation Detection (Textual Message Content Only) In Task A, which consist of binary detection of slur reclamation, participants are provided exclusively with the textual content of the social media posts. Approaches may rely solely on the supplied training data or may incorporate external resources, such as additional annotated datasets. In the latter case, the use of additional data had to be explicitly declared in participants' technical reports.

Task A is divided into three language-specific subtasks. Participants may choose to address one, two, or all three languages:
– **Subtask A1.** ***Italian:*** classification on Italian texts;
– **Subtask A2.** ***Spanish:*** classification on Spanish texts;
– **Subtask A3.** ***English:*** classification on English texts.

Task B - Contextual Slur Reclamation Detection Task B extends Task A by allowing participants to exploit information associated with the author's profiles, such as biographical descriptions, when available, in addition to the textual data. Task B includes two language-specific subtasks:
– **Subtask B1.** ***Italian:*** slur reclamation on Italian texts with contextual metadata;
– **Subtask B2.** ***Spanish:*** slur reclamation on Spanish texts with contextual metadata;

User biographical information was available only for Spanish and Italian. This information was not present in the English source resource.

Lang	User	Bio	Tweet	Reclamation
it	@user	Certo le circostanze non sono favorevoli. 🏳️ ✊ 🏳️‍🌈	In quanto disabile e frocia questi sono i miei PrideMonths. Ma vorrei anche dire che il giorno in cui nel manifesto di un evento lgbtqiatransfemminista verrà citato anche l'antiabilismo oltre ad anti sessismo/obtfobia/razzismo/specismo offro da bere	yes
it	@user	I veri partigiani furono i primi sovranisti! w la patria!	Ecco, adesso pensate all'iter o in affitto ed al male che fate al bambino branco di finocchi arcobaleno	no
es	@user	Me llaman feminista, roja y bollera.✊ 🏳️‍🌈 💜	Buenas tardes a rojos, feministas, republicanos, maricones, bolleras y demás LGTBI 🏳️‍🌈 #LGTBI #pride	yes
es	@user	I live for that energy!	Hace rato pasó una caravana de movimiento LGT...etc y algunos me vieron parado observando y me gritaron que yo también era marica. Órale, bien «respetuosas» estas personas que exigen respeto. #PrideMonth #Pride2022 #LGTBQ	no
en	N/A	N/A	I use the word tranny all the time...but that's only in reference to working on my cars. Transgendered, transvestite, and drag queen folk are too fabulous to have their descriptions abbreviated.	yes
en	N/A	N/A	Actually that's what s faggot is. Fag is just something that needs to be burnt.	no

Table 1
Reclaimed and not-reclaimed data examples from the multilingual dataset.

3. Dataset

For the purposes of this challenge, we constructed a multilingual dataset by combining existing publicly available resources in three languages: Italian, Spanish, and English. The data were collected from social media platforms between 2020 to 2022. Specifically, the dataset builds upon the following resources: TWITA, a large-scale Twitter (now X) corpus for Italian spanning over ten years [13]; the LGBTQI+ Dataset (2020-2022)_es, which contains Spanish tweets related to LGBTQ+ topics [14]; and the Reclaimed Hate Speech Dataset (RHSD) for English [11], from which only Twitter and Reddit content were retained.

To obtain a comparable multilingual collection, we applied a uniform multi-step filtering strategy across the three languages. First, we performed keyword-based filtering using homosexuality-related terms drawn from Hurtlex (OM category), a multilingual lexicon of hate speech [15]. This list was subsequently expanded through discussions within a multidisciplinary research team and consultations with the LGBTQ+ community (e.g. *frocio* in Italian, *faggot* in English, and *maricón* in Spanish).

In a second filtering phase, we refined the selection by targeting sentences with a higher likelihood of reclaimed language, identified through positive terms typically expressing pride and community belonging, such as *pride*, *rainbow*, *queer*, and related translations in each language. The complete list of keywords used for all three languages is reported in Appendix A, Table 7.

Table 1 shows some reclaimed and non-reclaimed examples of social media posts.

The dataset, named **MultiRECLAIM**, comprises a total of 4, 983 social media texts: 1, 811 in Italian,

1, 711 in English, and 1, 461 in Spanish. For the Italian and Spanish subsets, user biographical information is available, whereas such metadata is not provided in the English source resource. Given the highly context-dependent nature of slur reclamation, biographical information can support annotators by offering additional contextual cues and enabling considerations related to speaker legitimacy.

The **MultiRECLAIM** dataset was manually annotated by 165 native-speakers, comprising 30 Italian, 70 English, and 65 Spanish speakers. Each text was annotated by five annotators, with efforts made to balance annotator groups by gender and LGBTQ+ self-identification. Annotation was conducted through a dedicated interface, with participants recruited via Prolific crowd-sourcing platform[1] and direct recruitment. Sociodemographic information, including gender, sexual orientation, age, and country of birth, was collected during the process. Annotators were compensated above the U.S. federal minimum wage, estimating an average rate of £9.60 per hour. Participation was voluntary, and individuals could withdraw at any time without consequences.

Human annotators were engaged in a comprehensive annotation process based on a multi-label, fine-grained schema including the identification of offensiveness, irony, and stereotypes. For the purposes of the present task, we exclusively focus on the dimension of slur reclamation. Specifically, annotators were asked to answer the question: *Do you think the terms related to the LGBTQ+ context present in the sentence are used in a reclaimed way?*, thereby performing a binary classification task (yes/no).

The annotations were aggregated with majority voting. We report the label distribution in Table 2, alongside the Inter-Annotator Agreement calculated with Krippendorff's α.

Language	Positives	Negatives	Total	Krippendorff's α
Italian	346	1,465	1,811	0.6037
Spanish	221	1,240	1,461	0.3008
English	147	1,564	1,711	0.1553

Table 2
Label distribution and Krippendorff's α for each language.

During the MultiPRIDE challenge, each language-specific dataset was released separately and split into training and test sets. The training set (60%) was fully annotated and made available to participants under a Creative Commons Attribution 4.0 license, which defines terms for data usage and citation. The test set (40%) was released after the evaluation phase.

4. Evaluation

We provide a separate ranking for each subtask. This implies that teams that created a multilingual approach are still evaluated on each language separately. We also create a baseline for all subtasks.

4.1. Evaluation Metrics

We calculate Precision, Recall and, F1-score for all runs and the two binary labels.

$$Precision_{class} = \frac{\#correct_{class}}{\#assigned_{class}}$$

$$Recall_{class} = \frac{\#correct_{class}}{\#total_{class}}$$

$$F1_{class} = 2\frac{Precision_{class} * Recall_{class}}{Precision_{class} + Recall_{class}}$$

The final ranking solely based on **F1 macro-score**. As the tasks are all binary, the F1 macro is defined as:

$$F1_{macro} = (F1_{reclaimed} + F1_{not-reclaimed})/2$$

[1] https://www.prolific.com/

We also report recall, precision and F1-score for each label to better understand differences in performance between the various approaches.

Initially, we aimed to distinguish between *constrained* and *unconstrained* runs. However, persistent ambiguity around what qualifies as constrained made it difficult to apply this distinction consistently. This issue is largely driven by the fact that state-of-art Large Language Models are pre-trained on data that may incorporate information from annotated corpora, lexicons, and other resources. To avoid inconsistent or potentially misleading categorization, we ultimately chose not to label runs as constrained or unconstrained.

4.2. Baselines

For all subtasks, we used the same approach to create the baseline: a fine-tuned language specific BERT model. Specifically, the pre-trained models we chose are: BERT base Italian from MDZ Digital Library for Italian[2], BETO[3] [16] for Spanish, and RoBERTa base for English[4] [17]. We report all hyperparameters used for the baseline in Appendix B.

The only additional tweak we applied is to add class weights during loss calculation to contrast the severe class-imbalance between positive and negative labels. Also, for Task B, we include the bio by concatenating it to the text and apply different TYPE_TOKEN_IDS[5] to help the models distinguish the two separate information.

We voluntarily chose to have encoder-based baselines to compare them with LLMs which are decoder-based. Our aim is to effectively understand if there is a significant shift from the state-of-the-art performance of the LLMs for text classification or if their improvement is marginal compared to smaller and faster models.

5. Participants and Results

A total of 19 teams, comprising approximately 50 researchers, participated in the MultiPRIDE task. The majority of participants were affiliated with academic institutions, while a small number were independent researchers without formal institutional affiliations. The participating teams exhibited substantial geographical diversity, representing countries across Europe, Asia, and Latin America. European contributions originated from Italy, Spain, France, the Netherlands, and Romania, with Italy being the most represented country. Asian teams were based in Japan, Vietnam, and Iran, while Latin American participation was represented by Mexico.

Participants were allowed to submit systems in three languages—Italian, Spanish, and English—and, for each selected language, to compete in Subtask A, Subtask B, or both. For each language–task combination, teams could submit up to two runs. Table 3 provides an overview of the submissions across languages, tasks, and methods for all participating teams. One of the 19 teams, OUTLIERS, did not submit a final description report.

5.1. Methods

Fine-tuning LM Fine-tuning LM-based models was the dominant strategy, adopted by 17 out of 19 participating teams. Most approaches relied on multilingual or language-specific Transformer architectures. Several teams employed multilingual models such as XLM-RoBERTa or mDeBERTa (e.g., NAMDANG, LLANA, THE HATE BUSTERS), while others leveraved lightweight or parameter-efficient alternatives. Notably, I2C combined an LLM-based system fine-tuned via QLoRA on Qwen2.5-0.5B with a multilingual XLM-RoBERTa classifier. UNIBO-FICLIT adopted an ELECTRA-based encoder for binary

[2]https://huggingface.co/dbmdz/bert-base-italian-uncased
[3]https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[4]https://huggingface.co/FacebookAI/roberta-base
[5]This process is automatically done by the HuggingFace Tokenizer when passing a list shaped accordingly: [bio, text]

Team	Country	$\text{Task}_{Subtasks}$	LM Fine-tuning	Data augmentation	Multilingual	Decoder	Feature Extraction	Ensemble	Knowledge Injection
NetGuardAI [18]	Romania	$A_{1,2,3}$, $B_{1,2}$	✓						
I2C [19]	Spain	$A_{1,2,3}$	✓	✓	✓	✓			
NamDang [20]	Vietnam	$A_{1,2,3}$	✓	✓	✓				
Challenger [21]	Iran	$A_{1,2,3}$	✓	✓					
AIWizards [22]	Italy	$B_{1,2}$	✓				✓	✓	
MilaNLP [23]	Italy	A_1, B_1							✓
Ghavidel-Rajabi [24]	Italy, Iran	$A_{1,2,3}$	✓				✓	✓	
LlaNa [25]	Italy	A_1, B_1	✓	✓					
UniBO-FICLIT [26]	Italy	A_1, B_1	✓				✓		
HateItOff [27]	France, The Netherlands	$A_{1,2}$, $B_{1,2}$	✓				✓		
KIT-TIP-NLP [28]	Japan	$A_{1,2,3}$	✓	✓	✓				
Outliers*	Mexico	B_2	✓						
SaFe Tweets [29]	Italy	$A_{1,3}$, B_1	✓			✓		✓	
The Hate Busters [30]	Italy	$A_{1,2,3}$, $B_{1,2}$	✓		✓		✓		
GRUPPETTOZZO [31]	Italy	$A_{1,2,3}$, $B_{1,2}$	✓	✓	✓				
Kenji-Endo [32]	Italy	A_1, B_1	✓			✓			
INFOTEC [33]	Mexico	A_2	✓						
DataSummit [34]	Italy	$B_{1,2}$	✓				✓		
Avahi [35]	Mexico	$A_{1,2,3}$, $B_{1,2}$			✓	✓	✓		✓

Table 3
Overview of teams, tasks and methods. We marked with * the only team who did not submit a report.

classification, and SaFe Tweets proposed a two-stage pipeline using DeBERTaV3 with LoRA fine-tuning for initial filtering. Language-specific fine-tuning was also common: GRUPPETTOZZO employed umBERTo for Italian, BERT-base for English, and BETO for Spanish, while INFOTEC fine-tuned Spanish models specialized for Mexican and Latin American variants. Other teams incorporated additional signals, such as user biography information (NetGuardAI, AIWizards), sentiment and emotion features (DataSummit), or synthetic data generated via back-translation (KIT-TIP-NLP). Finally, while submitting their runs, Outliers only mentioned that they fine-tuned Robertuito (Spanish RoBERTa variant).

Data Augmentation Data augmentation was widely adopted, with 7 teams employing it primarily to mitigate class imbalance in the original dataset. I2C applied an augmentation technique based on Contextual Word Embeddings (CWE) exclusively to the training split. NamDang explored few-shot prompting with Large Language Models (LLMs) to generate synthetic samples for each target language. Several teams focused on translation-based augmentation strategies. Challenger implemented a comprehensive back-translation augmentation strategy on the training data using the Google Translate API, whereas LlaNa adopted a targeted cross-lingual approach by translating reclamation instances from English and Spanish datasets into Italian to strengthen the minority class. Similarly, KIT-TIP-NLP leveraged GPT-4o-mini for systemic multilingual back-translation, expanding the training corpus while preserving semantic content and class distribution.

In contrast, AIWizards introduced user-level soft labels indicating the likelihood of LGBTQ+ community membership, inferred from tweets and user biographies, to train a BERT-like model to predict community membership by considering latent representations associated with LGBTQ+ identity. Finally, GRUPPETTOZZO combined Full Dataset Augmentation and Balanced Augmentation applied to the

minority class strategies using MarianMT [36] and OPUS-MT-based [37, 38] back-translation pipeline to pipeline to generate semantically consistent paraphrases while introducing lexical and syntactic diversity.

Multilingual We want to underline the fact that part of the teams (I2C, NamDang, KIT-TIP-NLP, The Hate Busters, GRUPPETTOZZO, Avahi) created a single multilingual system to tackle more subtasks instead of creating a language-specific model. Thus, the difficulty of creating such system should be taken into consideration while looking at the results. Most of these approaches leveraged a multilingual BERT and opted to finetune it on all training splits for the different languages. Avahi created a single multilingual Knowledge Graph and a Retrieval Augmented Generation (RAG) system that took into account the language of the text to classify.

Decoder Four teams (I2C, SaFe Tweets, Kenji-Endo, Avahi) used a decoder model as the main backbone of their system. This was unexpected due to the attention given to generative models in the last years. Out of those five teams, none used a language specific model but rather leveraged LLMs that support multiple languages and have English as their main language in the pre-training. Only Avahi used a proprietary model, with others using open-weight models.

Feature Extraction One of the most popular approaches to the task is Feature Extraction. Most teams leveraged features related to the sentiment of the text. In fact, their underlying hypothesis was that negative sentiment correlated to offensive use of the terms and inversely correlated to reclamatory use of the slur. Other features that were used include emotion and syntactic information. We want to underline the efforts of Ghavidel-Rajabi [24] which defined new indicators that they call Paradox Score, *Sentiment-Hate Gap* and, *Reclamatory Intent Score*.

Ensemble Part of the teams relied on creating an ensemble of models where different streams were then unified to produce a single classification. SaFe Tweets leveraged an ensemble of lightweight models of different kinds: logistic-regression, SetFit, and a finetuned BERT-like model. Meanwhile, Ghavidel-Rajabi combined two different streams: the first based on embeddings of a LM and, the second based on the feature extraction they propose. AI Wizards proposed an interesting approach where they leverage two different encoder-based models. One finetuned on the task to detect the likelihood of an author being part of the LGBTQ+ community and the other finetuned on the detection of reclamation.

Knowledge Injection Two teams adopted knowledge injection strategies. Specifically, MilanNLP investigated retrieval-augmented approaches that leverage external knowledge derived from related hate speech corpora to assess whether such information can improve classification performance. To this end, they developed a retrieval-augmented pipeline using the HODI Subtask A corpus [39] as an external knowledge base. In contrast, Avahi employed a multilingual Knowledge Graph encoding reclaimed terms in Spanish and English for context injection. In particular, potential reclaimed terms identified via the Knowledge Graph were incorporated into the prompt, thereby providing Claude with explicit signals about the linguistic and semantic content of the message.

Pre-Training from Scratch A single team, Kenji-Endo pre-trained a BabyLM from scratch. This entails that they used a limited number of tokens in the pre-training. They propose two different approaches: (i) *Kenji-Endo Vanilla* a monolithic BabyLM and (ii) *Kendi-Endo MoE* a mixture of experts. As they participated in multiple EVALITA tasks, they finetuned their models on the specific task. For our task, they submitted a single run finetuning *Kendi-Endo Vanilla* for both Task A and B.

5.2. Results

We report the rankings of Task A and Task B in Table 4 and Table 5 respectively. The complete reports for all runs, which include Precision, Recall and, F1-score, can be found in Appendix C.

5.3. Subtask A

For Subtask A, the dominant approach consisted of transformer-based fine-tuning, either in monolingual or multilingual settings. XLM-RoBERTa emerged as the most widely used backbone (adopted by I2C, NamDang, Ghavidel–Rajabi, LlaNa, SaFe Tweets, The Hate Busters, and GRUPPETTOZZO), often in combination with multilingual training data. Several teams explored architectural comparisons, notably Ghavidel–Rajabi and NamDang, which evaluated both XLM-RoBERTa and mDeBERTa-v3, reaching the first position in Subtask A1 (Italian), with an F1-score of 0.8981, and the top-3 in Subtask A2-A3 (Spanish and English), respectively.

A number of teams relied on language-specific encoders, such as AlBERTo (MilaNLP), UmBERTo (GRUPPETTOZZO), BETO for Spanish (GRUPPETTOZZO), or ELECTRA (UniBO-FICLIT), for Italian, and BILMALAT and MEX_Large (INFOTEC) for Spanish. Notably, The Hate Busters and DataSummit applied Logistic Regression and SVM classifiers, where the first group obtained the best F1-score in Subtask A2 (Spanish). Lastly Avahi, who outperformed the Subtask A3 (English), adopted a multilingual Knowledge Graph and Claude 4.5 Haiku for final classification.

5.4. Subtask B

Subtask B generally followed similar modeling choices to Subtask A, but included a context integration step, most commonly by concatenating tweet text with user biography or profile information (strategy adopted by NetGuardAI, MilaNLP, GRUPPETTOZZO, AI Wizards, and more). Transformer fine-tuning remained the predominant approach, with XLM-RoBERTa and language-specific BERT variants again being the most frequently used architectures (I2C, LlaNa, SaFe Tweets, The Hate Busters). Within the mentioned groups, LlaNa obtained the highest F1-score of 0.9021 for Subtask B1 (Italian).

6. Discussion

6.1. Language Differences

Generally, we can observe that the Italian subtask proved to be easier than other languages with F1-scores hovering around the 0.9 threshold. We hypothesize that this is due to multiple factors that we mainly break down into:

- **Number of Positive Examples and Agreement**: the Italian sub-portion of the training corpus is the one with the highest number of *reclaimed* examples. Also, it is the language with the highest agreement between annotators, 0.6 according to Krippendorff's α. As both the IAA and model performance is higher for Italian, we can hypothesize that reclamation signals were easier to understand, leading to an easier task.
- **Slurs Distribution**: analyzing the data, we found that Italian had big discrepancies in use for certain slurs. In fact, specific slurs were almost always used in a *reclamatory* manner, meanwhile, others were almost always used in a *non-reclamatory* manner. We report the data regarding each slur for each language in Table 6. In fact, we can see that the word *trans* made up a huge percentage of negative examples. Meanwhile, *forcio*, *forcia* and, *forci* are exclusively used in positive examples, rendering the label easier. We can observe similar patterns of distributions for Spanish and English but not to this extreme.

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) Results for Subtask A1 (Italian)

Team	Run	F1-score
The Hate Busters	2	0.7776
NamDang	2	0.7707
NamDang	1	0.7590
infotec	2	0.7569
GRUPPETTOZZO	2	0.7506
HateItOff	1	0.7370
KIT-TIP-NLP	1	0.7312
GRUPPETTOZZO	1	0.7289
Avahi	1	0.7117
I2C	2	0.7105
The Hate Busters	1	0.7098
infotec	1	0.7034
KIT-TIP-NLP	2	0.7016
baseline	1	0.7000
Ghavidel-Rajabi	1	0.6960
Challenger	1	0.6774
NetGuardAI	1	0.6476
HateItOff	2	0.6442
I2C	1	0.4892

(b) Results for Subtask A2 (Spanish)

Team	Run	F1-score
Avahi	1	0.6416
SaFe Tweets	1	0.6329
NamDang	1	0.6305
Avahi	2	0.6225
Challenger	1	0.6204
GRUPPETTOZZO	2	0.5979
NamDang	2	0.5901
baseline	1	0.5760
I2C	2	0.5708
The Hate Busters	2	0.5500
NetGuardAI	1	0.5476
Ghavidel-Rajabi	1	0.5460
The Hate Busters	1	0.5291
GRUPPETTOZZO	1	0.5285
Ghavidel-Rajabi	2	0.5275
I2C	1	0.4945
KIT-TIP-NLP	2	0.4545
KIT-TIP-NLP	1	0.3973

(c) Results for Subtask A3 (English)

Table 4
F1 Macro-Average Scores for Task A

Team	Run	F1-score	Δ with Task A
LlaNa	1	0.9021	+0.0186
baseline	1	0.8981	+0.0250
GRUPPETTOZZO	1	0.8979	+0.0145
The Hate Busters	2	0.8827	+0.0324
MilaNLP	1	0.8827	-0.0132
GRUPPETTOZZO	2	0.8681	-0.0054
UniBO-FICLIT	1	0.8644	-0.0063
MilaNLP	2	0.8641	+0.0335
AIWizards	1	0.8564	-
AIWizards	2	0.8549	-
HateItOff	2	0.8462	-0.0122
The Hate Busters	1	0.8324	-0.0020
HateItOff	1	0.8319	-0.0490
SaFe Tweets	1	0.8164	-0.0731
Avahi	1	0.8031	+0.0127
DataSummit	1	0.7734	-
Kenji-Endo	1	0.7489	-0.0871
NetGuardAI	1	0.7405	-0.1132

(a) Results for Subtask B1 (Italian)

Team	Run	F1-score	Δ with Task A
Outliers	1	0.7378	-
GRUPPETTOZZO	1	0.7304	+0.0015
The Hate Busters	1	0.7196	+0.0097
The Hate Busters	2	0.7195	-0.0581
Avahi	1	0.7006	-0.0111
HateItOff	1	0.7005	-0.0365
baseline	1	0.6966	-0.0034
AIWizards	2	0.6950	-
GRUPPETTOZZO	2	0.6857	-0.0649
HateItOff	2	0.6733	+0.0291
DataSummit	1	0.6361	-
AIWizards	1	0.5935	-
NetGuardAI	1	0.4927	-0.1549

(b) Results for Subtask B2 (Spanish)

Table 5
F1 Macro-Average Scores for Task B. For teams that have submitted runs for both Task A and B, we include the difference between the two F1 Macro.

Slur	Pos.	Neg.
checca	2	4
culattone	0	3
finocchi(o)	10	14
finocchietto	0	1
forci(o)	32	0
forci	25	0
forcia	4	0
froci(o)	43	105
frocia/e	12	3
lella	8	0
ricchione/i	14	15
travestita/e	0	2
travestiti/o	1	22
trans	5	451

(a) Label Distribution for Slurs in Italian

Slur	Pos.	Neg.
bollera(s)	10	12
bollos	0	1
gay(s)	12	89
marica(s)	24	31
maricón(es)	38	97
queer	5	242
trans	9	185
travestis	6	2

(b) Label Distribution for Slurs in Spanish

Slur	Pos.	Neg.
bitch	0	52
drag	2	11
fag	5	22
faggot(s)	48	422
gay	13	143
queen	0	7
queer	17	104
whore	1	6
slut	0	5
sissy	0	2
trans	4	24

(c) Label Distribution for Slurs in English

Table 6
Label Distribution for Slurs in the Test Set; **Pos** indicates the *reclaimed* instances containing the slur; **Neg.** indicates the *not reclaimed* instances containing the slur.

6.2. Baseline Performance and Encoder Models

As explained in Section 4.2, we deliberately chose encoder models as our baseline because they were the most popular systems in the most recent EVALITA campaigns. In fact, we hypothesized that the major shift of the NLP community towards decoder models would also be felt in our task, thus, the baseline could prove how much LLMs affected the state-of-the-art in classification tasks. Surprisingly, most of the runs we received do not actually incorporate LLMs as the main building block of their system.

The importance of encoder-models is also reflected in their performance. Our proposed baselines proved to be quite challenging in almost every subtask, achieving almost always average position and being outperformed by 0.07 in the worst case. Surprisingly, the baseline came in second place for Subtask B1, with a single team overcoming it by a small margin.

This leads us to wonder about the performance of decoder-models. In fact, only one of the five subtasks has been won by a run that relied on LLMs for classification. Interestingly, said task is A3, which is the only English task and the proposed system relied on a proprietary-model (Claude 4.5 Haiku). This can raise a few questions about decoder models in general:

- Are open-weights LLMs actually reliable for these types of classification tasks?
- Have efforts to develop language-specific decoder models been successful? [40]
- Do proprietary models still favor English despite their efforts towards multilinguality?

We realized that these are massive research questions that we cannot answer by relying on our single task, but we hope that our results can contribute to the broader discussion regarding these topics. In our case, we found that: open-weights models still do not perform reliably; we did not receive any run using a language-specific decoder; and that the only win achieved by a LLM was in the English subtask.

6.3. Use of Contextual Information

The only difference between Task A and Task B is related to the possibility of using contextual information about the user, i.e. their social media biographies. Reasonably, it could be expected that the use of additional information about the identity of the author could help the classification systems. In fact, a previous EVALITA task, Sardistance [41] observed major improvements in the contextual subtask compared to textual-only subtask. Crucially, we expected this to happen in our challenge as reclamation is a phenomenon which heavily depends on the identity of the writer.

For clarity, in Table 5, we report the difference in F1-score between Task A and B for each run submitted for both tasks. In more than half cases, we record a decrease in performance by systems using additional information. In cases where systems improve, the increase in performance is not very substantial (a maximum of +0.03). Interestingly, both LLaNa and GRUPPETTOZZO which were two of the best performing teams in Subtask B1 and B2, utilized data augmentation in two different manners. The former included translations of the English subset to Italian and, the latter used back translations.

To better understand the difficulties posed by the subtask, we report the causes listed by the teams which recorded the biggest drops in performance. NetGuardAI hypothesized that the root of the problem is that biographies often included generic text (e.g., "Music lover", "Madrid") which acted as noise. This is also stated by SaFe Tweets who concatenated bio and texts and found that the signal of the reclamation markers was diluted by the bios. An interesting case is Kenji-Endo, the team proposed a BabyLM, meaning that they performed a pre-training with limited tokens. They noted that the lack of emojis in their data sources negatively affected the model performance in Subtask B1. In fact, they state "the mean byte-fallback ratio is 0.1149 for biographies compared to 0.0134 for tweet text, and 34.3% of biography entries contain at least one such token (versus 9.9% for tweet text)" [32].

7. Conclusion and Future Works

This paper introduces MultiPRIDE, the first shared task on slur reclamation in a multilingual context. The task aims to identify instances in which slurs lose their derogatory intent and are instead used in positive contexts to express identity and belonging. We analyzed the submissions of 19 participating teams and found that slur reclamation constitutes a highly subjective and complex phenomenon, strongly influenced by cultural and linguistic factors. Moreover, system performance varied substantially across languages, suggesting that models struggle to generalize reclamation patterns beyond language-specific cues. This work also presents several limitations, primarily related to the dataset, including the imbalance between reclaimed and non-reclaimed instances. The limited availability of reclaimed examples contributed to reduced performance and higher variance across systems; however, this imbalance reflects the real-world distribution of the slur reclamation phenomenon in online spaces, as it is a highly specific and relatively rare phenomenon. Additionally, the two-step filtering process may introduce a strong lexical bias, as reclaimed instances are more likely to appear in contexts that explicitly reference pride-related concepts. This creates a risk that models rely on surface-level cues rather than deeper signals of reclamation intent. Against this backdrop, future research should focus on extending the dataset to other languages in order to improve coverage and robustness. Furthermore, the integration of culturally grounded features represents a valuable direction, as cultural context can significantly shape the perception of slur reclamation even within the same language. Additionally, we plan to extend the investigation of slur usage to other communities beyond LGBTQ+ community by incorporating additional terms and considering different intersectional dimensions (such as ethnicity and geographic provenance). This will allow us both to enlarge the dataset and to analyze whether models exhibit performance variations when handling slurs reclaimed by different communities. Finally, another line of future work concerns the use of generative AI to produce synthetic reclaimed data, which could help mitigate data scarcity and support more balanced data distribution and experiments.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (version 5.1), Claude (Sonnet 4.5) in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

The work of Paolo Rosso was in the framework of the ANNOTATE-MULTI2 research project (Grant PID2024-156022OB-C32) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. The work of Viviana Patti has been partially supported by the "HARMONIA" project - M4-C2, I1.3 Partenariati Estesi - Cascade Call - FAIR - CUP C63C22000770006 - PE PE0000013 under the NextGenerationEU programme.

References

[1] 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.

[2] E. W. Pamungkas, V. Basile, V. Patti, Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6237–6246. URL: https://aclanthology.org/2020.lrec-1.765.

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

[4] E. Holgate, I. Cachola, D. Preoţiuc-Pietro, J. J. Li, Why swear? analyzing and inferring the intentions of vulgar expressions, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, ACL, Brussels, Belgium, 2018, pp. 4405–4414.

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

[6] B. Cepollaro, D. L. de Sa, The successes of reclamation, Synthese 202 (2023) 205.

[7] B. Cepollaro, P. Labinaz, M. Fasoli, F. Ervas, T. Piazza, M. Croce, T. Numerico, F. Panzeri, G. Tuzet, F. Domaneschi, L. De Vita, S. Di Paola, M. G. Rossi, J. Menichetti, F. Macagno, Social network e comunicazione, Sistemi Intelligenti 2019 (2019) 385–632.

[8] T. Gillespie, Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media, Yale University Press, 2018.

[9] N. Strossen, Hate: Why we should resist it with free speech, not censorship, Oxford University Press, 2018.

[10] B. Cepollaro, M. Lepoutre, R. M. Simpson, Counterspeech, Philosophy Compass 18 (2023) e12890. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12890.

[11] 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, 2024, pp. 241–249.

[12] L. Draetta, C. Ferrando, M. Cuccarini, L. James, V. Patti, ReCLAIM Project: Exploring Italian Slurs Reappropriation with Large Language Models, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 335–342.

[13] V. Basile, M. Lai, M. Sanguinetti, Long-term social media data collection at the university of turin, in: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), CEUR-WS, 2018, pp. 1–6.

[14] J. Mata, E. Gualda, A dataset of Spanish tweets on people and communities LGBTQI+ during the COVID-19 pandemic 2020-2022 [LGBTQI+ Dataset 2020-2022_es], A dataset of Spanish tweets on people and communities LGBTQI+ during the COVID-19 pandemic 2020-2022 [LGBTQI+ Dataset 2020-2022_es] (2025).

[15] 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), CEUR Workshop Proceedings, Turin, Italy, 2018, pp. 52–57.

[16] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: Practical ML for Developing Countries Workshop at ICLR 2020, 2020, pp. 1–10.

[17] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[18] N. C. J. Le Pollotec, E. S. Apostol, C. O. Truică, NetGuardAI at MultiPRIDE: Multilingual Detection of Reclaimed Language in the MultiPRIDE 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.

[19] L. Vázquez-Ramos, J. Mata-Vázquez, V. Pachón-Álvarez, I2C at MultiPRIDE: Transformers and Generative LLMs for Multilingual Reclamation Classification, 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.

[20] N. Dang, T. Dang, V. T. Kiet, NamDang at MultiPRIDE: Multilingual Classification of Reclaimed Language in LGBTQ+ Discourse using Transformer-based Models, 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.

[21] H. B. A. Tekanlou, M. Bakhtiyarzadeh, J. Razmara, Challenger at MultiPRIDE: Is It Hate Speech or Reclaimed?, 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.

[22] L. Tedeschini, M. Fasulo, AIWizards at MultiPRIDE: A Hierarchical Approach to Slur Reclamation Detection, 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.

[23] I. Crescenzi, A. Muti, D. Nozza, MilaNLP at MultiPRIDE: Evaluating Lexical, Transformer, and Retrieval-Augmented Models for Reclaimed Language, 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.

[24] H. Rajabi, F. Ghavidel, K. Ghahremani, Ghavidel-Rajabi at MultiPRIDE 2026: Identity, Toxicity, or Complexity? A Language-Specific Feature Selection Approach to Reclamatory Intent Detection, 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.

[25] A. Mercurio, G. Talluto, I. Siragusa, R. Pirrone, LlaNa at MultiPRIDE: A fine-tuning approach with cross-lingual augmentation, 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.

[26] S. Casazza, UniBO-FICLIT at MultiPRIDE: Fine-Tuning an ELECTRA-Based Model for the Detection of Italian Reclaimed Slurs, 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.

[27] G. Damo, N. B. Ocampo, HateItOff at MultiPRIDE: Linguistic and Sentiment Cues in Reclaimed LGBTQ+ Slur Detection, 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.

[28] B. G. HB, M. Ptaszynski, R. Melendez, J. Eronen, KIT-TIP-NLP at MultiPRIDE: Continual Learning with Multilingual Foundation Model, 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.

[29] S. Visconti, F. Manzi, SaFeTweets at MultiPRIDE: From Multi-Model Ensembles to Expert LLMs for Reappropriation Detection, 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.

[30] A. Ciminelli, G. Corvino, C. Gentili, M. Viviani, The Hate Busters at MultiPRIDE: Automatic Identification of Reappropriated Slurs in Multilingual Social Media, 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.

[31] F. Traina, A. Santoro, G. Greco, I. Siragusa, R. Pirrone, GRUPPETTOZZO at MultiPRIDE: Detecting LGBTQ+ Reclamatory Intent via Context-Aware Transformers, 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.

[32] Scozzaro, Calogero Jerik and Rinaldi, Matteo and Mittone, Gianluca and Stranisci, Marco Antonio, Kenji-Endo: a BabyLM @EVALITA, 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.

[33] J. Gleaves, G. Ruiz, INFOTEC-NLP at MultiPRIDE: Spanish Reclaimed Language Classification, 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.

[34] F. Dingeo, M. Viviani, DataSummit at MultiPRIDE: 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.

[35] T. Alcántara, O. Garcia-Vazquez, J. A. Torres-León, M. Cardoso-Moreno, D. Jimenez, L. Moreno-Mendieta, AVAHI at MultiPRIDE: Multilingual Reclaimed Language Detection via Knowledge Graphs and Retrieval-Augmented Generation, 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.

[36] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, A. Birch, Marian: Fast neural machine translation in C++, in: F. Liu, T. Solorio (Eds.), Proceedings of ACL 2018, System Demonstrations, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 116–121. URL: https://aclanthology.org/P18-4020/. doi:10.18653/v1/P18-4020.

[37] J. Tiedemann, S. Thottingal, OPUS-MT–Building open translation services for the World, in: Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, 2020, pp. 479–480.

[38] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato, Y. Scherrer, R. Vázquez, S. Virpioja, Democratizing neural machine translation with OPUS-MT, Language Resources and Evaluation 58 (2024) 713–755.

[39] 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), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 1–8. URL: https://ceur-ws.org/Vol-3473/paper26.pdf.

[40] B. Magnini, M. Madeddu, M. Resta, R. Zanoli, M. Cimmino, P. Albano, V. Patti, A Leaderboard for Benchmarking LLMs on Italian, in: C. Bosco, E. Jezek, M. Polignano, M. Sanguinetti (Eds.), Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 636–646. URL: https://aclanthology.org/2025.clicit-1.61/.

[41] A. T. Cignarella, M. Lai, C. Bosco, V. Patti, P. Rosso, Sardistance @ EVALITA2020: overview of the task on stance detection in italian tweets, in: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 1–10.

A. Dataset Keywords Filtering

Table 7 reports the full set of keywords used in the first and second filtering stages for each language. These keywords were employed to filter the original datasets and to select the texts included in our final datasets.

Language	Stage	Keywords	Final output
Italian [13]	1° filtering	checca, culattone, finocchio, frocio, travestito, trans, ricchione, stesso sesso, legge164, lella, forcio, forci,...	1,890
	2° filtering	pride, lgbt, orgoglio, queer, arcobaleno, diritti, cultura gay, storia gay, lotta, fobia	
English [11]	1° filtering	fag, bitch, gay, puss, queer, drag, whore, queen, slut, sissy, slay	1,711
	2° filtering	pride, lgbt, queer, proud, rainbow, phobia, rights, human, fight	
Spanish [14]	1° filtering	bollera, bolleras, bollos, marica, maricón, maricon, maricones, queer	1,461
	2° filtering	pride, lgbt, orgullo, arcoíris, fobia, derecho, cultura gay, historia gay, lucha	

Table 7
Keyword-based filtering procedure for dataset construction.

B. Baseline Hyperparameters

We report the hyperparameters for the baseline models in Table 8. We did not conduct any hyperparameter search but rather used standard-values for detection tasks for similar domains.

Hyperparameter	Value
Training Epochs	3
Batch Size	16
Learning Rate	$2e^{-5}$
Seq. Length Task A	240
Seq. Length Task B	480
Seed	13

Table 8
Hyperparameter used for the baseline approaches.

C. Complete Results

We report the complete results which include Precision, Recall and, F1 for all labels in Table 9 (for Subtask A1), Table 10 (for Subtask A2), Table 11 (for Subtask A3), Table 12 (for Subtask B1) and, Table 13 (for Subtask B2).

Team	Run	Neg. Label			Pos. Label			Macro Avg.		
		Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Ghavidel-Rajabi	1	0.9517	0.9744	**0.9629**	0.8800	0.7914	**0.8333**	0.9158	0.8829	**0.8981**
MilaNLP	1	0.9561	0.9659	0.9610	0.8496	0.8129	0.8309	0.9029	0.8894	0.8959
Ghavidel-Rajabi	2	0.9441	0.9795	0.9615	0.8974	0.7554	0.8203	0.9208	0.8675	0.8909
SaFe Tweets	1	0.9412	0.9829	0.9616	0.9115	0.7410	0.8175	0.9263	0.8620	0.8895
LlaNa	1	0.9452	0.9710	0.9579	0.8618	0.7626	0.8092	0.9035	0.8668	0.8835
GRUPPETTOZZO	1	0.9409	0.9778	0.9590	0.8879	0.7410	0.8078	0.9144	0.8594	0.8834
Challenger	1	0.9451	0.9693	0.9570	0.8548	0.7626	0.8061	0.9000	0.8659	0.8816
HateItOff	1	0.9436	0.9710	0.9571	0.8607	0.7554	0.8046	0.9021	0.8632	0.8809
GRUPPETTOZZO	2	0.9521	0.9505	0.9513	0.7929	0.7986	0.7957	0.8725	0.8745	0.8735
baseline	1	0.9294	0.9881	0.9578	0.9314	0.6835	0.7884	0.9304	0.8358	0.8731
UniBO-FICLIT	1	0.9319	0.9812	0.9559	0.8981	0.6978	0.7854	0.9150	0.8395	0.8707
NamDang	2	0.9411	0.9539	0.9475	0.7939	0.7482	0.7704	0.8675	0.8511	0.8589
HateItOff	2	0.9194	**0.9932**	0.9549	**0.9565**	0.6331	0.7619	**0.9380**	0.8131	0.8584
NetGuardAI	1	0.9482	0.9369	0.9425	0.7466	0.7842	0.7649	0.8474	0.8605	0.8537
The Hate Busters	2	**0.9757**	0.8925	0.9323	0.6667	**0.9065**	0.7683	0.8212	**0.8995**	0.8503
I2C	2	0.9303	0.9573	0.9437	0.7951	0.6978	0.7433	0.8627	0.8276	0.8435
NamDang	1	0.9372	0.9420	0.9396	0.7500	0.7338	0.7418	0.8436	0.8379	0.8407
Kenji-Endo	1	0.9272	0.9556	0.9412	0.7851	0.6835	0.7308	0.8561	0.8195	0.8360
The Hate Busters	1	0.9383	0.9334	0.9358	0.7254	0.7410	0.7331	0.8318	0.8372	0.8345
MilaNLP	2	0.9295	0.9454	0.9374	0.7519	0.6978	0.7239	0.8407	0.8216	0.8306
KIT-TIP-NLP	2	0.9123	0.9590	0.9351	0.7798	0.6115	0.6855	0.8461	0.7853	0.8103
Avahi	1	0.9252	0.9078	0.9165	0.6400	0.6906	0.6644	0.7826	0.7992	0.7904
KIT-TIP-NLP	1	0.9012	0.9334	0.9170	0.6695	0.5683	0.6148	0.7853	0.7509	0.7659
I2C	1	0.9020	0.6911	0.7826	0.3442	0.6835	0.4578	0.6231	0.6873	0.6202

Table 9
Detailed Performance for Subtask A1 (Italian). Best scores are highlighted in bold.

Team	Run	Neg. Label			Pos. Label			Macro Avg.		
		Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
The Hate Busters	2	0.9287	0.9437	0.9361	0.6500	0.5909	**0.6190**	0.7894	0.7673	**0.7776**
NamDang	2	0.9237	0.9497	**0.9365**	0.6622	0.5568	0.6049	0.7929	0.7533	0.7707
NamDang	1	0.9229	0.9396	0.9312	0.6203	0.5568	0.5868	0.7716	0.7482	0.7590
infotec	2	0.9372	0.9014	0.9190	0.5421	0.6591	0.5949	0.7396	0.7802	0.7569
GRUPPETTOZZO	2	0.9421	0.8833	0.9117	0.5126	0.6932	0.5894	0.7273	0.7882	0.7506
HateItOff	1	0.9086	0.9598	0.9335	0.6667	0.4545	0.5405	0.7876	0.7072	0.7370
KIT-TIP-NLP	1	0.9501	0.8431	0.8934	0.4583	0.7500	0.5690	0.7042	0.7965	0.7312
GRUPPETTOZZO	1	0.9149	0.9296	0.9222	0.5625	0.5114	0.5357	0.7387	0.7205	0.7289
Avahi	1	**0.9794**	0.7666	0.8600	0.4082	**0.9091**	0.5634	0.6938	**0.8378**	0.7117
I2C	2	0.9038	0.9457	0.9243	0.5846	0.4318	0.4967	0.7442	0.6887	0.7105
The Hate Busters	1	0.9287	0.8652	0.8958	0.4508	0.6250	0.5238	0.6898	0.7451	0.7098
infotec	1	0.8948	0.9759	0.9336	**0.7209**	0.3523	0.4733	**0.8079**	0.6641	0.7034
KIT-TIP-NLP	2	0.9606	0.7847	0.8638	0.4022	0.8182	0.5393	0.6814	0.8014	0.7016
baseline	1	0.9562	0.7907	0.8656	0.4023	0.7955	0.5344	0.6793	0.7931	0.7000
Ghavidel-Rajabi	1	0.8953	0.9638	0.9283	0.6400	0.3636	0.4638	0.7677	0.6637	0.6960
Challenger	1	0.9707	0.7344	0.8362	0.3684	0.8750	0.5185	0.6696	0.8047	0.6774
NetGuardAI	1	0.9394	0.7485	0.8331	0.3386	0.7273	0.4621	0.6390	0.7379	0.6476
HateItOff	2	0.8812	**0.9698**	0.9234	0.6053	0.2614	0.3651	0.7432	0.6156	0.6442
I2C	1	0.8504	0.7203	0.7800	0.1524	0.2841	0.1984	0.5014	0.5022	0.4892

Table 10
Detailed Performance for Substask A2 (Spanish). Best scores are highlighted in bold.

Team	Run	Neg. Label			Pos. Label			Macro Avg.		
		Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Avahi	1	0.9401	0.9281	0.9341	0.3284	0.3729	**0.3492**	0.6342	**0.6505**	**0.6416**
SaFe Tweets	1	0.9385	0.9265	0.9325	0.3134	0.3559	0.3333	0.6260	0.6412	0.6329
NamDang	1	0.9351	0.9441	0.9396	**0.3396**	0.3051	0.3214	**0.6374**	0.6246	0.6305
Avahi	2	0.9368	0.9233	0.9300	0.2941	0.3390	0.3150	0.6155	0.6312	0.6225
Challenger	1	0.9356	0.9281	0.9318	0.2969	0.3220	0.3089	0.6162	0.6251	0.6204
GRUPPETTOZZO	2	0.9351	0.8978	0.9161	0.2381	0.3390	0.2797	0.5866	0.6184	0.5979
NamDang	2	0.9270	0.9537	0.9402	0.2927	0.2034	0.2400	0.6099	0.5785	0.5901
baseline	1	**0.9402**	0.8291	0.8812	0.1955	0.4407	0.2708	0.5679	0.6349	0.5760
I2C	2	0.9255	0.9329	0.9292	0.2222	0.2034	0.2124	0.5739	0.5681	0.5708
The Hate Busters	2	0.9203	**0.9776**	**0.9481**	0.3000	0.1017	0.1519	0.6102	0.5397	0.5500
NetGuardAI	1	0.9240	0.8930	0.9082	0.1625	0.2203	0.1871	0.5432	0.5567	0.5476
Ghavidel-Rajabi	1	0.9199	0.9728	0.9457	0.2609	0.1017	0.1463	0.5904	0.5373	0.5460
The Hate Busters	1	0.9188	0.9217	0.9203	0.1404	0.1356	0.1379	0.5296	0.5287	0.5291
GRUPPETTOZZO	1	0.9179	0.9649	0.9408	0.1852	0.0847	0.1163	0.5516	0.5248	0.5285
Ghavidel-Rajabi	2	0.9178	0.9808	0.9483	0.2500	0.0678	0.1067	0.5839	0.5243	0.5275
I2C	1	0.9155	0.8131	0.8613	0.0930	0.2034	0.1277	0.5042	0.5082	0.4945
KIT-TIP-NLP	2	0.9303	0.5974	0.7276	0.1095	0.5254	0.1813	0.5199	0.5614	0.4545
KIT-TIP-NLP	1	0.9325	0.4633	0.6190	0.1016	**0.6441**	0.1755	0.5170	0.5537	0.3973

Table 11
Detailed Performance for Subtask A3 (English). Best scores are highlighted in bold.

Team	Run	Neg. Label			Pos. Label			Macro Avg.		
		Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
LlaNa	1	**0.9625**	0.9625	0.9625	0.8417	**0.8417**	**0.8417**	0.9021	**0.9021**	**0.9021**
baseline	1	0.9517	0.9744	**0.9629**	0.8800	0.7914	0.8333	0.9158	0.8829	0.8981
GRUPPETTOZZO	1	0.9562	0.9676	0.9618	0.8561	0.8129	0.8339	0.9061	0.8903	0.8979
The Hate Busters	2	0.9394	0.9795	0.9591	0.8947	0.7338	0.8063	0.9171	0.8567	0.8827
MilaNLP	1	0.9394	0.9795	0.9591	0.8947	0.7338	0.8063	0.9171	0.8567	0.8827
GRUPPETTOZZO	2	0.9278	**0.9863**	0.9562	0.9216	0.6763	0.7801	0.9247	0.8313	0.8681
UniBO-FICLIT	1	0.9343	0.9710	0.9523	0.8534	0.7122	0.7765	0.8939	0.8416	0.8644
MilaNLP	2	0.9371	0.9659	0.9513	0.8347	0.7266	0.7769	0.8859	0.8462	0.8641
AIWizards	1	0.9367	0.9590	0.9477	0.8080	0.7266	0.7652	0.8723	0.8428	0.8564
AIWizards	2	0.9310	0.9676	0.9490	0.8362	0.6978	0.7608	0.8836	0.8327	0.8549
HateItOff	2	0.9187	0.9829	0.9497	0.8980	0.6331	0.7426	0.9083	0.8080	0.8462
The Hate Busters	1	0.9296	0.9471	0.9383	0.7578	0.6978	0.7266	0.8437	0.8225	0.8324
HateItOff	1	0.9216	0.9625	0.9416	0.8053	0.6547	0.7222	0.8634	0.8086	0.8319
SaFe Tweets	1	0.9009	0.9932	0.9448	**0.9494**	0.5396	0.6881	**0.9251**	0.7664	0.8164
Avahi	1	0.9422	0.8908	0.9158	0.6257	0.7698	0.6903	0.7840	0.8303	0.8031
DataSummit	1	0.9117	0.9164	0.9140	0.6397	0.6259	0.6327	0.7757	0.7711	0.7734
Kenji-Endo	1	0.9350	0.8345	0.8819	0.5198	0.7554	0.6158	0.7274	0.7949	0.7489
NetGuardAI	1	0.9280	0.8362	0.8797	0.5127	0.7266	0.6012	0.7204	0.7814	0.7405

Table 12
Detailed Performance for Subtask B1 (Italian). Best scores are highlighted in bold.

Team	Run	Neg. Label			Pos. Label			Macro Avg.		
		Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Outliers	1	0.9022	0.9839	**0.9413**	0.8140	0.3977	0.5344	**0.8581**	0.6908	**0.7378**
GRUPPETTOZZO	1	0.9194	0.9175	0.9184	0.5393	0.5455	0.5424	0.7293	0.7315	0.7304
The Hate Busters	1	0.9329	0.8672	0.8989	0.4634	0.6477	0.5403	0.6982	0.7575	0.7196
The Hate Busters	2	0.9021	0.9638	0.9319	0.6667	0.4091	0.5070	0.7844	0.6864	0.7195
Avahi	1	**0.9766**	0.7565	0.8526	0.3950	0.8977	**0.5486**	0.6858	**0.8271**	0.7006
HateItOff	1	0.8992	0.9517	0.9247	0.5932	0.3977	0.4762	0.7462	0.6747	0.7005
baseline	1	0.9498	0.7988	0.8678	0.4012	0.7614	0.5255	0.6755	0.7801	0.6966
AIWizards	2	0.9119	0.8954	0.9036	0.4639	0.5114	0.4865	0.6879	0.7034	0.6950
GRUPPETTOZZO	2	0.8933	0.9598	0.9253	0.6078	0.3523	0.4460	0.7506	0.6560	0.6857
HateItOff	2	0.8879	0.9718	0.9280	0.6585	0.3068	0.4186	0.7732	0.6393	0.6733
DataSummit	1	0.9051	0.8249	0.8632	0.3409	0.5114	0.4091	0.6230	0.6682	0.6361
AIWizards	1	0.9130	0.7183	0.8041	0.2784	0.6136	0.3830	0.5957	0.6660	0.5935
NetGuardAI	1	0.8537	**0.9980**	0.9202	0.7500	0.0341	0.0652	0.8019	0.5160	0.4927

Table 13
Detailed Performance for Subtask B2 (Spanish). Best scores are highlighted in bold.