I2C At MultiPRIDE: Transformers And Generative LLMs For Multilingual Reclamation Classification

Laura Vázquez-Ramos[1,*,†], Jacinto Mata-Vázquez[1,†] and Victoria Pachón-Álvarez[1,†]

[1]I2C Research Group, Information Technology Department, Universidad de Huelva, Fuerzas Armadas Avenue, 21007, Huelva, Spain

Abstract

Online hate and abusive language targeting LGBTQ+ people are increasingly common on social media. Such content can spread prejudice, intimidate users, and limit participation in online spaces. Automatic detection is therefore important to support moderation and to reduce exposure to harmful content. This paper presents our submissions to MultiPRIDE 2026 Task 1 for Italian, Spanish, and English, comparing two modelling paradigms: an encoder-based Transformer sequence classifier and a generative large language model adapted to the task through supervised instruction tuning. Results are reported with macro-averaged F1 to account for class imbalance. Considering the best-performing run per subtask, our system achieves macro-F1 = 0.571 (EN), 0.711 (ES), and 0.843 (IT) on the official test sets, showing the strongest performance in Italian and consistent improvements across languages. These results suggest that discriminative multilingual Transformers remain a robust choice for intent-sensitive reclamation detection under limited context.

Content Warning: This paper contains examples of explicit and offensive language, including hate speech and slurs, presented for scientific analysis and documentation purposes.

Keywords

Natural Language Processing, LGBTQ+ Hate Speech, Abusive Language Detection, Transformers, Large Language Models

1. Motivation

Online texts often include terms related to LGBTQ+ identity. Sometimes they are used as insults. Other times they are used by the community itself, as self-labels or jokes. This second case is known as reclaimed language: words that were once offensive but are re-used with a positive or in-group meaning [1].

Distinguishing between reclamatory and offensive usage is important for both moderation and discourse analysis. If reclaimed language is not recognised, legitimate messages may be flagged as harmful, resulting in the disproportionate removal of LGBTQ+ voices. Conversely, if hostile uses are misinterpreted as reclamatory, targeted abuse may remain undetected. For this reason, the central challenge is not simply the presence of specific terms, but the intent with which they are used.

This problem is challenging because intent is highly context-dependent. It may be conveyed indirectly through irony, euphemisms, or figurative language, and its interpretation can vary across languages and communities. Under text-only conditions—without conversational history or user-level information—key pragmatic cues are missing, increasing ambiguity and making reliable classification difficult.

Within this context, the MultiPRIDE 2026 Task 1 [2] provides a multilingual benchmark for evaluating systems that decide whether an LGBTQ+ related term in a sentence is used with reclamatory intent or not, across Italian, Spanish, and English. The goal of this work is to contribute to this evaluation by developing and analysing two complementary modelling paradigms—Transformer-based sequence classification and generative LLM-based inference—and by examining the factors that most strongly affect performance across languages.

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.

laura.vazquez@dti.uhu.es (L. Vázquez-Ramos); mata@uhu.es (J. Mata-Vázquez); vpachon@uhu.es (V. Pachón-Álvarez)

0009-0001-3974-8943 (L. Vázquez-Ramos); 0000-0001-5329-9622 (J. Mata-Vázquez); 0000-0003-0697-4044 (V. Pachón-Álvarez)

id	text	label
it_1231	La destra Italiana pur di non dire che loro odiano gli omosessua...rola Identità di Genere.....ovviamente a la #Salvini #omnibusla7	0
it_1713	"Presupporre che tutti i bisessuali non sono mai attratti da per... e genderqueer che scelgono di etichettarsi come bisessuali" URL	0
it_1770	io devo sapere chi è l'intern forcio che si occupa della musica a cortesie per gli ospiti URL	1
it_1155	A volte mi ricordo di quanto sia bello essere lella e non poterlo dire con fierezza a nessuno	1

Table 1
Examples from the Italian dataset (Subtask A1).

2. Definition of the Task

This work addresses MultiPRIDE 2026 Task 1, which focuses on reclaimed language in the LGBTQ+ community. The goal is to study how terms related to LGBTQ+ identity are used in online text and to highlight the challenges that arise when a term can be either derogatory or used with a reclamatory, in-group intent. Systems are expected to rely on cues that can be expressed in the text itself (e.g., slurs, denigratory words, self-labeling, figurative language), and, in general, to capture the intended meaning behind the use of these terms.

Task 1 is formulated as a binary classification problem. Given a sentence that contains a term related to an LGBTQ+ context, the system must predict whether the term is used with reclamatory intent or non-reclamatory intent.

We participated in the text-only setting (Task A – Textual Content), where the input consists solely of the message text and no user-profile information is provided. Participation was carried out in all three language-specific subtasks:

- **Subtask A1 (Italian)**: classification on Italian texts.
- **Subtask A2 (Spanish)**: classification on Spanish texts.
- **Subtask A3 (English)**: classification on English texts.

3. Dataset

The organisers released one dataset per language track. Each instance includes an identifier (*id*), the message text (*text*), and a binary label (*label*) indicating whether the LGBTQ+-related term is used with reclamatory intent (1) or non-reclamatory intent (0). A language tag (*lang*) is also provided. For Italian and Spanish, an additional field (*bio*) is included; however, it was not used in our experiments because we participated in the text-only setting.

Three monolingual datasets were provided, corresponding to Italian (Subtask A1), Spanish (Subtask A2), and English (Subtask A3). The Italian dataset contains 1,086 instances, with 879 (80.94%) labelled as class 0 and 207 (19.06%) as class 1. The Spanish dataset contains 876 instances, with 743 (84.82%) instances in class 0 and 133 (15.18%) in class 1. The English dataset contains 1,026 instances and is the most imbalanced split, with 938 (91.42%) instances labelled as class 0 and 88 (8.58%) as class 1.

In addition to the three monolingual datasets, a multilingual training set was created by merging Italian, Spanish, and English into a single pool to support multilingual training. This setup was treated as an unconstrained approach, since an additional derived dataset was constructed for training purposes. Before balancing, the merged dataset contained 2,988 instances in total, with 2,560 instances in class 0 and 428 instances in class 1.

Illustrative examples for each language are reported in Tables 1, 2, and 3, respectively. Table 4 shows the class distribution of the three monolingual datasets and the merged multilingual dataset, reporting the number of instances for class 0 and class 1 in each case.

id	text	label
es_1850	28 de Junio - Día Internacional del Orgullo LGTBI. Viva el amor Libre!!!#lgtb #lgtbiq #lgtbi #amorlibre #loveislove #orgullogay #pride #...#lovewins #orgullo #respeto #lucha #homosexual #doblajespain URL	0
es_773	@USER no me gusta la Montero, por su apoyo a la ley Trans, ignor... es un acoso y derribo a una mujer que lucha por sus hijos.. URL	0
es_1196	¡¡¡Ya estamos de MARIcañas y BOLLOpintxos!!!#Orgullo2021 #OrgulloRioja2021 #OrgulloLGTBI URL	1
es_1884	Soy Queer.#Queer #Orgullo #Pride #Orgullo2020 #OrgulloLGTBI #LGTBIQ #trescantos URL	1

Table 2
Examples from the Spanish dataset (Subtask A2).

id	text	label
en_1021	I've never heard anyone use the word "faggot" and not sound like an insecure 13 year old.	0
en_1496	So you don't see the slighest problem of somebody calling other people "faggots"?	0
en_1312	And to be fair, getting triggered by slurs is just stupid to begin with. Yeah I'm a faggot, yeah I'm gay. I'm mo...l because it has more syllables and thus takes more time to say.	1
en_1108	Tbh I don't care if anyone says it as long as they say it jokingly. Also happy I'm not the only one who calls themselves faggot	1

Table 3
Examples from the English dataset (Subtask A3).

Dataset	Class 0	Class 1	Total
Italian (A1)	879	207	1086
Spanish (A2)	743	133	876
English (A3)	938	88	1026
Multilingual (IT+ES+EN)	2560	428	2988

Table 4
Class distribution for the three monolingual datasets and the merged multilingual dataset (before balancing).

4. Description of the System

This section describes the methodology adopted, which is structured into two main stages: balancing and splitting the datasets and training models.

4.1. Balancing and Splitting the Datasets

For development, an internal 80/20 split was applied to obtain training and evaluation partitions. Since the original datasets are strongly imbalanced, a data augmentation technique based on Contextual Word Embeddings (CWE) was applied only on the training split [3]. Specifically, we utilized a BERT-based architecture, the bert-base-uncased model [4]. This method entails substituting selected tokens with semantically similar alternatives derived from the model's contextual representations, thereby increasing the lexical diversity of the corpus while preserving the semantic integrity of the original instances. In this case, part of the majority class (label 0) was removed and new examples of the minority class (label 1) were generated until reaching 1,500 instances per label. As a result, the balanced training sets contain 3,000 instances each.

Before applying CWE-based balancing, each dataset was split using an 80/20 scheme. The 80% portion was used for model development, while the remaining 20% was kept as an unseen test set for the final evaluation. Table 5 reports the class distribution for each subtask before and after applying

Dataset	Label	Before balancing (Train)	After balancing (Train)
A1 (Italian)	0 (non-reclamatory)	703	1500
	1 (reclamatory)	165	1500
A2 (Spanish)	0 (non-reclamatory)	594	1500
	1 (reclamatory)	106	1500
A3 (English)	0 (non-reclamatory)	750	1500
	1 (reclamatory)	70	1500
Multilingual (IT+ES+EN)	0 (non-reclamatory)	2048	1500
	1 (reclamatory)	342	1500

Table 5
Label distribution in the training split before and after CWE-based balancing (balancing applied only to Train; the Test split was kept unchanged).

Dataset	Train	Valid	Test
A1 (Italian)	2400	600	218
A2 (Spanish)	2400	600	176
A3 (English)	2400	600	206
Multilingual (IT+ES+EN)	2400	600	598

Table 6
Number of instances in the training, validation, and test splits. CWE-based balancing was applied only to the initial Train split (80%).

data augmentation on the development portion only. The development split was then further divided using another 80/20 split to obtain the final training and validation sets. The validation set was used for model selection and hyperparameter tuning, whereas the test set was reserved exclusively for reporting results. Table 6 reports the number of instances assigned to the training and test partitions for each subtask.

4.2. Training models

Since the shared task allowed two runs per subtask, two complementary modelling approaches were developed for each language (Italian, Spanish, and English). A first run was based on a Large Language Model (LLM) [5] adapted to the task through supervised instruction tuning (QLoRA), where the model was guided to predict whether the target term was used with reclamatory intent or not. A second run relied on a Transformer model [6] using the provided training data, aiming to learn task-specific decision boundaries from labelled examples. This two-run setup was designed to compare a standard fine-tuning approach against a prompt-based alternative under the same evaluation protocol, and to assess how each paradigm behaves better in each subtask.

4.2.1. Run 1: QLoRA Fine-tuning

For the first run submitted for each subtask, an LLM-based approach was adopted. Models were fine-tuned with QLoRA [7], which enables parameter-efficient training by combining low-rank adapters with quantised weights, making it feasible to adapt LLMs under limited computational resources while preserving most of the base model's knowledge. The main backbone used across experiments was *Qwen2.5-0.5B* [8], selected as a lightweight instruction-capable model that can be fine-tuned efficiently and deployed consistently across languages. In addition, a different backbone was used for Spanish: *teknium/OpenHermes-2.5-Mistral-7B*[1]. This choice was motivated by the strong availability of Spanish-friendly instruction tuning in the Mistral/OpenHermes family and by empirical development results, where Spanish performance benefited from a larger, instruction-focused model. In practice, this design

[1] https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B

Hyperparameter	Search space	Best value
Learning rate	{3e-5, 5e-5, 8e-5, 1e-4}	1e-4
Train batch size	{2, 4, 8}	2
Eval batch size	{16}	16
Gradient accumulation steps	{1, 2, 4}	4
LR scheduler	{linear, cosine}	linear
Warmup ratio	{0.03, 0.05, 0.1}	0.03
Weight decay	{0.0, 0.01}	0.01
Label smoothing	{0.0, 0.02}	0.0

Table 7
LLM hyperparameter search space and best configuration selected on the validation split.

provided a balance between efficiency and coverage (Qwen2.5-0.5B) and capacity for Spanish-specific instruction following (OpenHermes-2.5-Mistral-7B) [9].

A hyperparameter study was carried out for the LLM-based classifier in order to obtain a stable configuration before producing the final runs. The study was implemented with the Hugging Face Trainer using Optuna[2] as the optimisation backend. For each trial, the model was fine-tuned and evaluated on the validation split, and macro-F1 was used as the optimisation target. In total, 15 trials were run. The search space covered the main training hyperparameters that most strongly affect performance and convergence. The best hyperparameter configuration selected by this procedure is summarised in Table 7.

The multilingual LLMs were fine-tuned exclusively on the multilingual merged dataset described in Section 3. Training instances were formatted using an instruction-style prompt. The base prompt was: *"Eres un asistente lingüístico. Clasifica si el uso del término LGTBQ+ es reapropiativo (1) o no (0). No des la etiqueta en el texto, solo aprende del patrón.\n\nTexto:\n"*

For each training instance, the prompt was dynamically translated to match the language of the input text, while preserving the same instructional structure and semantics across languages. This approach ensures cross-lingual consistency and frames the task as supervised instruction tuning, where the model is fine-tuned to learn a direct mapping between the input text and the binary classification label indicating reappropriative or non-reappropriative usage.

In contrast, language-specific fine-tuning was only applied in the Spanish setting, where a Spanish-oriented LLM was trained using the Spanish monolingual dataset and a Spanish prompt. No language-specific LLMs were trained for English or Italian.

4.2.2. Run 2: Multilingual Fine-tuning with XLM-RoBERTa

For the second run submitted for each subtask, a Transformer-based classifier was trained by fine-tuning XLM-RoBERTa-base [10] on the multilingual merged dataset only. Using a single multilingual training set was a deliberate choice. First, the task definition is identical across languages, and a shared training pool increases the diversity of surface forms (slurs, self-labels, and figurative uses) that the model is exposed to. Secondly, a unified multilingual model avoids maintaining separate language-specific fine-tuned checkpoints and encourages cross-lingual transfer, where patterns learned in one language can support predictions in another.

XLM-RoBERTa-base was selected because it is a widely used multilingual encoder with strong performance on cross-lingual text classification and stable fine-tuning behaviour across many languages, including Italian, Spanish, and English. Its pretraining on large-scale multilingual corpora provides robust shared representations, which makes it well suited for the single-model, multilingual training setup adopted in this run. The hyperparameters were kept fixed to ensure comparability between models, using a *learning rate* of 3e-5, *batch size* of 16 and training for 10 epochs with *Early Stopping*. These values were chosen because they are widely used and have been shown to yield stable performance when fine-tuning Transformer-based models for text classification [4, 11].

[2] https://optuna.org/

5. Evaluation Measures

Models were assessed on the validation data using standard metrics for binary classification derived from the confusion matrix [12]. In particular, *precision*, *recall*, *accuracy*, and *F1-score* were computed. In addition, the *Area Under the ROC Curve (AUC)* was reported as a threshold-independent measure of how well the system separates the two classes [13]. Among these metrics, *accuracy*, *F1-score*, and *AUC* were taken as the main indicators of performance.

Accuracy is the simplest evaluation measure in binary classification and corresponds to the fraction of instances that are correctly predicted. Although it is easy to interpret, it can be misleading when classes are imbalanced or when the minority class is the one of interest. For this reason, it is reported together with F1-score and AUC. Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives.

The F1-score combines precision and recall into a single value and is especially informative when the positive class is under-represented. It captures the trade-off between correctly identifying positive cases and avoiding false alarms, and is therefore well suited for hate and abuse detection tasks where errors on the minority class are critical. It is defined as:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

The AUC complements these point estimates by summarising performance across all possible decision thresholds. It measures the probability that a randomly selected positive instance is ranked above a randomly selected negative one, providing a broader view of discriminative capacity than accuracy alone [14]. Formally, it can be expressed as:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{3}$$

where TPR and FPR are the true positive rate and false positive rate along the ROC curve [15].

In hate-speech-related tasks, where class distributions are typically skewed and abusive intent may be expressed indirectly, relying on a single metric can hide important failure modes. Reporting F1-score together with AUC makes it possible to evaluate both the balance between precision and recall and the system's ability to distinguish between classes under varying thresholds, while accuracy is retained as an additional reference point for overall correctness.

In addition, the final decision criterion was aligned with the official shared-task evaluation protocol. F1-score was used as the primary metric for model selection and comparison, while the remaining measures (accuracy, precision, recall, and AUC) were reported as complementary indicators to better characterise system behaviour.

6. Results

This section reports the experimental results obtained for Task 1 across the three language subtasks. Results are presented in two parts, reflecting the two stages of the workflow. First, we summarise the scores observed during the training phase, where models were developed and compared using the balanced datasets and the train/validation split described in the previous sections. Next, we present the evaluation-phase results, corresponding to the official runs submitted to the shared task and their scores on the hidden test sets.

Run	Model / Training setup	Accuracy	Macro-F1	AUC
Run 1 (LLM + QLoRA)	OpenHermes-2.5-Mistral-7B (Spanish)	0.7483	0.7475	0.7483
	Qwen2.5-0.5B (Multilingual)	0.7483	0.7454	0.7483
Run 2 (Transformer)	XLM-RoBERTa-base (Multilingual)	0.9117	0.9115	0.9100

Table 8
Training-phase results on the validation split.

Subtask	Run 1 (LLM)			Run 2 (Transformer)		
	Macro-P	Macro-R	Macro-F1	Macro-P	Macro-R	Macro-F1
A1 (Italian)	0.6231	0.6873	0.6202	0.8627	0.8276	0.8435
A2 (Spanish)	0.5014	0.5022	0.4892	0.7442	0.6887	0.7105
A3 (English)	0.5042	0.5082	0.4945	0.5739	0.5681	0.5708

Table 9
Official evaluation results on the shared-task test sets for both runs (macro-averaged precision, recall, and F1).

Subtask	Run 1 (LLM)						Run 2 (Transformer)					
	$\mathbf{P}_0$	$\mathbf{R}_0$	$\mathbf{F1}_0$	$\mathbf{P}_1$	$\mathbf{R}_1$	$\mathbf{F1}_1$	$\mathbf{P}_0$	$\mathbf{R}_0$	$\mathbf{F1}_0$	$\mathbf{P}_1$	$\mathbf{R}_1$	$\mathbf{F1}_1$
A1 (Italian)	0.9213	0.8589	0.8890	0.3249	0.5156	0.3980	0.9676	0.9324	0.9496	0.7577	0.7228	0.7398
A2 (Spanish)	0.9146	0.9083	0.9114	0.0882	0.0962	0.0920	0.9322	0.9309	0.9316	0.5563	0.4465	0.4954
A3 (English)	0.9204	0.9072	0.9138	0.0880	0.1091	0.0974	0.9349	0.9244	0.9296	0.2129	0.2117	0.2123

Table 10
Per-class precision (P), recall (R), and F1 for both runs on the shared-task test sets. Class 1 corresponds to reclamatory intent and class 0 to non-reclamatory intent.

6.1. Training-phase Results

In this subsection, the results obtained during the training phase are reported for the models developed for each subtask and for each run. These scores were computed on the validation split and are used to summarise how the different modelling choices performed during development, before producing the final submissions.

Table 8 summarises the results obtained during the training phase for the two submitted runs. Overall, the multilingual XLM-RoBERTa model achieves the strongest scores on the development split. For the LLM-based run, different backbones were explored: while Qwen2.5-0.5B was adopted as the default choice due to its efficiency and stable behaviour, the Spanish track was trained with OpenHermes-2.5-Mistral-7B because it yielded slightly better validation performance in our experiments, and was therefore selected for the final Spanish submission.

6.2. Evaluation-phase Results

This subsection reports the official results obtained on the shared-task test sets for the two submitted runs. For each language subtask, we report the macro-averaged precision, recall, and F1-score.

Table 9 summarises overall behaviour through macro-averaged precision, recall, and F1, making the cross-language comparison straightforward and directly reflecting the competition ranking metric. In this view, the Transformer-based run achieves the best macro-F1 in all three subtasks, with the strongest performance in Italian and Spanish and a smaller, yet consistent, improvement in English.

Table 10 helps explain these trends by reporting performance separately for the two classes. Across languages, high scores for class 0 indicate that both approaches recognise the majority *non-reclamatory* category reliably. The main source of difficulty is class 1 (*reclamatory intent*), where precision and recall are substantially lower. Importantly, the Transformer run yields a clear improvement for class 1 compared to the LLM run, especially in Italian and Spanish, which largely accounts for the gains observed in macro-F1. This confirms that overall performance in this task is driven primarily by the ability to detect the minority class, rather than by differences in handling the majority class.

These results suggest that multilingual fine-tuning with XLM-RoBERTa provides a robust solution for this task under the competition evaluation setting.

7. Discussion

7.1. Qualitative Error Analysis

To complement the quantitative results, a qualitative error analysis was conducted using the official error logs released for Italian, Spanish, and English. Each file contains 20 misclassified instances, balanced across the two labels. Although the sample is small, it is useful to spot repeated patterns and to understand why the models fail beyond what is shown by global scores. A clear trend is that most errors are linked to the minority class, i.e., reclaimed usage. This is expected in a setting where the same term can act as an insult in one context and as an in-group marker in another.

Several recurring sources of confusion were observed across languages. A first issue is metalinguistic mention. Many texts do not use the term to attack or to reclaim it, but to talk about it: users quote someone else, discuss whether the word should be considered offensive, or criticise its usage. In these cases, the models seem to react strongly to the presence of the term itself. As a result, cues that signal stance are sometimes missed, such as negation, reported speech, or explicit condemnation.

A second issue is irony and humour, which are common in reclaimed language. Users may apply the term to themselves or to friends in a playful way, often with informal tone and profanity. This mixture can be difficult to interpret automatically. Even when the text contains signals of affiliation, the model may still lean towards the non-reclamatory class because aggressive wording is present.

A third source of error is topic mixing. Some messages combine identity-related terms with political debate, activism, or discussions about rights and legislation. These contexts are often polarised and emotionally loaded. The term can appear as a neutral reference, as part of a slogan, or as a weaponised label. When only the message text is available, it is not always clear which intent is dominant.

Finally, surface variation also plays a role. Creative spellings, wordplay, compounding, and morphological variants can reduce robustness, especially when these forms are rare in the training data. In practice, small changes in form may hide the lexical cues that the model has learned to rely on.

Table 11 shows representative examples (five per language) that illustrate these patterns. The same types of ambiguity appear in all subtasks, but with different emphasis. In English, many errors are linked to quotation and to mention-versus-use ambiguity. In Spanish, mistakes are often found in Pride or activism posts with dense hashtag context, where reclaimed usage is mixed with slogans and community tags. In Italian, errors include explicit in-group framing as well as politically oriented discussion, where intent is difficult to infer from text alone.

Overall, the error analysis suggests that further gains will likely require better modelling of stance and clearer separation between mentioning a term and using it directly. Targeted augmentation could also help, especially if it focuses on irony, self-reference, activism contexts, and political discourse. Under an unconstrained setting, additional context (e.g., conversational cues) could reduce ambiguity, but the text-only condition makes these cases particularly challenging.

7.2. Conclusions and Future works

This work explored two complementary approaches for identifying reclamatory intent in LGBTQ+-related terms across Italian, Spanish, and English: a QLoRA fine-tuned LLM run and a multilingual Transformer run based on XLM-RoBERTa-base. Overall, the Transformer submission achieved the strongest performance across all subtasks, while the LLM-based approach provided a competitive baseline and offered a flexible framework for rapid experimentation. Across languages, performance was mainly limited by the minority class, confirming that reclaimed usage remains difficult to detect reliably under a text-only setting.

The qualitative analysis suggests that most remaining errors are linked to ambiguous pragmatic phenomena, including metalinguistic mentions (talking about a term rather than using it), quotation

Subtask	id	Gold_label	Text
A1 (Italian)	it_169	1	chiamate ancora Louis froc...CHÉ ANCHE IO SONO UNA FROCIA DI MERDA MA NON TOCCATEMI LOUIS)
A1 (Italian)	it_1585	1	Che Libreria Adelphi sia chiusa nel weekend è un affronto a noi ricchione lavoratrici
A1 (Italian)	it_44	1	Sono in pari con #Euphoria, grazie a questa queen nei panni di una lella #Rules URL
A1 (Italian)	it_913	0	qui siamo tutti dolci e im...rché avevo bevuto 2 birre. ma chi dovrei essere? gesù cristo?
A1 (Italian)	it_489	0	Domani in Spagna verrà app...allentano l'approvazione del #DDLZan. Invidio gli spagnoli.
A2 (Spanish)	es_422	1	Hoy un poco más mariconeada, pues #OrgulloLGTBI URL
A2 (Spanish)	es_500	1	FELIZ DÍA MARICONES, y a t...basura, que os atropelle el coche de amaia #OrgulloLGTBI URL
A2 (Spanish)	es_1189	1	mariquísimo #NiQueer #NiTERF — #Orgullo2021 #Orgullo #OrgulloLGTBI URL
A2 (Spanish)	es_1199	0	Un año fuera del closet ...ER #pride #Pride2021 #Orgullo2021 #Orgullo #OrgulloLGTBI
A2 (Spanish)	es_1484	0	Hay q joderse Resulta q s... Trans y una Transgenero, yo ya lo flipo. Hágaselo mirar. URL
A3 (English)	en_1527	1	> "We do not hire faggots" ...only the retarded.
A3 (English)	en_1565	1	YES! Queer and faggot (bu... against me in a negative way and I just can't get over that.
A3 (English)	en_1106	1	I'm Canadian. I use queer... it light years away from "faggot" in terms of offensiveness.
A3 (English)	en_420	0	Ally I know My family is proud of me faggot can you say the same?
A3 (English)	en_1872	0	I've been out for a few weeks and I'm dreading the day when I get called a faggot

Table 11
Representative misclassified examples (5 per subtask) selected from the error logs.

and reported speech, ironic self-reference, and politically polarised contexts. These cases often require modelling stance and intent beyond surface lexical cues, which explains why high accuracy on the majority class can coexist with substantially lower scores for reclamatory intent.

Future improvements could follow three directions. First, training data could be enriched with more targeted examples covering the observed error patterns (e.g., pride/activism posts, ironic self-labeling, and quoted mentions), ideally with augmentation strategies that preserve intent while varying surface forms. Secondly, the models could be improved by teaching them to tell the difference between mentioning a term and using it. In many posts, the word appears because people quote it, talk about whether it is offensive, or report what someone said. In these cases, the term is present, but it is not being used with a direct intent. This could be addressed by adding extra training signals or prompt steps that make the system look for stance cues (e.g., negation, quotation, reported speech) and decide whether the speaker is condemning the term, discussing it, or using it in an in-group way before predicting the final label. Finally, under an unconstrained setting, incorporating additional context (e.g., conversational thread information or user-level signals when available) could reduce ambiguity and improve classification of borderline cases, especially for the reclamatory class.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for grammar and spelling checks. All outputs were subsequently reviewed and edited by the authors, who take full responsibility for the content of the publication.

References

[1] M.-C. Schreuder, Safe spaces, agency, and resistance: A metasynthesis of lgbtq language use, Journal of LGBT Youth 18 (2021) 256–272.

[2] C. Ferrando, L. Draetta, M. Madeddu, M. Sosto, V. Patti, P. Rosso, C. Bosco, J. Mata, E. Gualda, Multipride at evalita 2026: Overview of the multilingual automatic detection of slur reclamation in the lgbtq+ context task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[3] E. Sezerer, S. Tekir, A survey on neural word embeddings, 2021. URL: https://arxiv.org/abs/2110.01804. doi:10.48550/arXiv.2110.01804. arXiv:2110.01804.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[5] A. Zubiaga, Natural language processing in the era of large language models, 2024.

[6] S. Islam, H. Elmekki, A. Elsebai, J. Bentahar, N. Drawel, G. Rjoub, W. Pedrycz, A comprehensive survey on applications of transformers for deep learning tasks, Expert Systems with Applications 241 (2024) 122666.

[7] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in neural information processing systems 36 (2023) 10088–10115.

[8] Q. Team, et al., Qwen2 technical report, arXiv preprint arXiv:2407.10671 2 (2024).

[9] M. Vangeli, Large language models as advanced data preprocessors: Transforming unstructured text into fine-tuning datasets, 2024.

[10] X. Ou, H. Li, Ynu@ dravidian-codemix-fire2020: Xlm-roberta for multi-language sentiment analysis., in: FIRE (Working Notes), 2020, pp. 560–565.

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

[12] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, arXiv preprint arXiv:2008.05756 (2020). doi:10.48550/arXiv.2008.05756.

[13] J. V. Carter, J. Pan, S. N. Rai, S. Galandiuk, Roc-ing along: Evaluation and interpretation of receiver operating characteristic curves, Surgery 159 (2016) 1638–1645. doi:10.1016/j.surg.2015.12.029.

[14] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, BMC genomics 21 (2020) 6. doi:10.1186/s12864-019-6413-7.

[15] A. P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, Pattern recognition 30 (1997) 1145–1159. doi:10.1016/S0031-3203(96)00142-2.