The Hate Busters at MultiPRIDE: Automatic Identification of Reappropriated Slurs in Multilingual LGBTQ+ Discourse

Aurora Ciminelli[†], Giulia Corvino[†], Camilla Gentili[†] and Marco Viviani[*]

Università degli Studi di Milano-Bicocca, Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo) – Edificio U14 (ABACUS), Viale Sarca, 336 – 20126 Milan, Italy

Abstract

This paper presents the methodology developed for the MultiPRIDE shared task by the Hate Busters team, which focuses on automatically identifying reappropriative intent in multilingual social media messages containing LGBTQ+ related slurs. Reappropriation is a nuanced sociolinguistic phenomenon in which historically derogatory terms are deliberately reclaimed with positive or community-affirming meanings. To model this behavior, we work with the multilingual dataset provided for the MultiPRIDE task, including texts in Italian, Spanish, and English. Our approach integrates a comprehensive methodological pipeline, encompassing preprocessing of noisy social media text, the construction of traditional sparse lexical representations such as Bag-of-Words and TF–IDF, and contextual Transformer-based embeddings from XLM-RoBERTa, followed by training linear classifiers and fine-tuned Transformer models. The goal is to systematically compare how different modelling paradigms capture the subtle and context-dependent characteristics of reappropriative language.

Warning: This paper contains examples of explicitly offensive content.

Keywords

Reappropriative intent, Semantic reclamation, LGBTQ+, Natural Language Processing, Text Classification

1. Introduction

Semantic reclamation is a linguistic phenomenon in which terms that have historically carried derogatory or offensive meanings are reappropriated by the communities they originally targeted, acquiring new, neutral, or empowering connotations [1, 2, 3]. While reclaimed terms may retain their offensive potential in certain contexts, their meaning can shift significantly depending on the speaker, audience, and communicative intent. This phenomenon is particularly evident in the use of slurs within the LGBTQ+ community, where words traditionally employed as insults are often used with a reclaiming intent to express identity, solidarity, or empowerment [4]. In online environments, especially on social media platforms, distinguishing between derogatory and reclaiming uses of such language becomes a critical challenge, as surface-level lexical cues are often insufficient to capture the underlying intent. It is within this context that MultiPRIDE at EVALITA 2026 [5] is situated. The MultiPRIDE task [6] focuses on identifying reclaimed slurs in Italian, Spanish, and English within content related to the LGBTQ+ community.

The approach proposed by the Hate Busters team within MultiPRIDE integrates a comprehensive methodological pipeline designed to tackle the challenges posed by the identification of reclaimed slurs in social media text. The pipeline begins with preprocessing techniques specifically tailored to handle the noisy and informal nature of user-generated content, including normalization, tokenization, and cleaning of textual data. Following preprocessing, we construct both traditional sparse lexical representations, such as Bag-of-Words and TF–IDF, and dense contextual embeddings derived from Transformer-based models, in particular XLM-RoBERTa. These representations are then employed

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.

a.ciminelli1@campus.unimib.it (A. Ciminelli); g.corvino2@campus.unimib.it (G. Corvino); c.gentili3@campus.unimib.it (C. Gentili); marco.viviani@unimib.it (M. Viviani)

https://ikr3.disco.unimib.it/people/marco-viviani/ (M. Viviani)

0000-0002-2274-9050 (M. Viviani)

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

to train linear classifiers as well as fine-tuned Transformer models, allowing us to investigate the relative effectiveness of different modelling paradigms in capturing the subtle and context-dependent characteristics of reclamatory language. Our main contributions are the integration of a robust and flexible pipeline capable of handling multilingual social media data, the systematic comparison of sparse and dense text representations for the detection of reclaimed slurs, and the evaluation of both classical and state-of-the-art Transformer-based models. This approach highlights how different techniques can capture the nuanced meanings that arise in semantic reclamation and provides a structured methodology for future work on intent-sensitive classification tasks in the context of LGBTQ+ language use.

2. The MultiPRIDE Dataset and Tasks

This section provides information on the dataset released by MultiPRIDE, an analysis of these data, and an overview of the tasks proposed.

2.1. Description of the Dataset

The dataset provided by MultiPRIDE is a multilingual corpus composed of textual social media content in Italian, Spanish, and English, aggregated from heterogeneous data sources. As specified in the official MultiPRIDE guidelines [5], the data originate from:

- Italian: the *TWITA* collection, a large-scale corpus of Italian tweets [7];
- Spanish: the *LGBTQI+ Dataset 2020–2022*, consisting of tweets related to LGBTQ+ topics [8];
- English: a composite dataset drawn from Twitter, Reddit, and TV series transcripts focusing on reclaimed and potentially derogatory terminology [9].

All texts were collected from publicly available sources between 2020 and 2022.

2.2. Preliminary Analysis of the Data

The initial datasets, provided by the organizers of the MultiPRIDE challenge, were harmonized through a uniform filtering pipeline applied across the three languages. A first selection step was conducted via keyword-based filtering, using terms related to homosexuality and derogatory expressions derived from the *Hurtlex* lexicon (e.g., fag, gay, bitch, etc.). To further refine the corpus and increase the likelihood of retrieving instances of reclaimed language, a second filtering step targeted positively connoted and community-oriented terms, such as pride, queer, LGBT, and rainbow.

After the automated filtering phase, all candidate texts were manually annotated with a binary label indicating whether an LGBTQ+-related term was used with **reappropriative intent** (*yes*) or not (*no*).

The initial set of data includes 1,811 observations for Italian, 1,711 for English, and 1,461 for Spanish. An initial exploratory analysis of the data showed that, in all languages, the distribution of labels is not balanced, with non-reclaimed instances representing the majority class. As illustrated in Figure 1, which reports the label distribution computed by merging the three languages' training sets, the dataset is strongly skewed toward the non-reclaimed category. This imbalance motivates the adoption of evaluation metrics that are robust to skewed label distributions.

We also examined the distribution of message lengths, measured in both tokens and characters. As expected for social media content, the majority of texts are relatively short; however, the presence of a long tail of longer messages was particularly informative for subsequent modelling choices. As shown in Figure 2, which reports the token-length distribution computed over the merged multilingual dataset, most texts fall between 20 and 60 tokens, with a clear concentration around the average length of 35.6 tokens. The maximum observed length is 148 tokens, resulting in a heavily right-skewed distribution. Since the choice of the maximum sequence length hyperparameter directly affects memory usage and the risk of losing information through truncation, this analysis was essential for setting an appropriate threshold. Therefore, the analysis of these properties supported different preprocessing decisions described later.

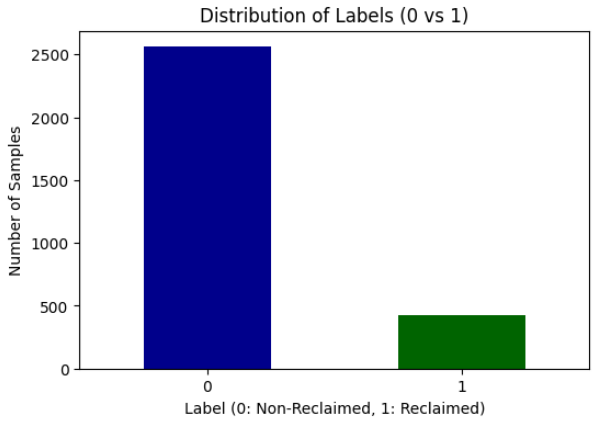

Figure 1: *Distribution of labels in the merged multilingual training set (Italian, English, Spanish).*

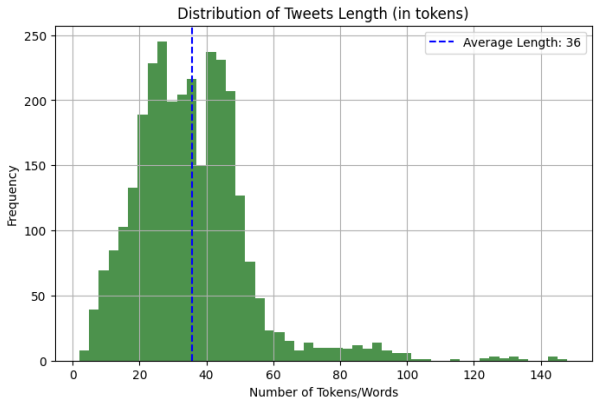

Figure 2: *Distribution of tweet lengths (in tokens). The dashed line marks the mean length.*

The training data are distributed in three distinct CSV files, one for each language. All files share a common structure defined by the following fields:

- **id**: Unique identifier of the message;
- **text**: Textual content of the message;
- **label**: Binary label (*yes* = reclaimed, *no* = not reclaimed);
- **bio**: Author biography, when available;
- **lang**: Language of the message.

In addition, contextual information about the author's profile is available only for the Italian and Spanish datasets, where the *bio* field contains short biographical descriptions extracted from the user's public profile. An inspection of missing values confirms this distribution: the bio field is absent for 100% of English instances, while missing rates are comparatively low for Italian (≈8%) and Spanish (≈6%) entries in the training data, resulting in an overall missing proportion of ≈39% across languages. Similar patterns are observed in the test set. This additional information is specifically relevant to Task B, which permits the use of contextual metadata alongside the textual content in order to improve the classification of reappropriative intent.

A preview of the dataset structure is reported in Figure 3, which displays a subset of Italian instances and illustrates the format and content of the fields.

Finally, we also examined the lexical distribution of the corpus through word-cloud visualizations, distinguishing between the two labels (see Figure 4 and Figure 5). The word cloud associated with Label 0 highlights the prevalence of denigratory terms as well as high-frequency but non-informative tokens (e.g., user, url). On the other hand, the word cloud for Label 1 reflects the linguistic complexity

id	text	bio	label	lang
it_1779	hanno fatto un discorso a tavola fra la differenza tra “frocio e ricchione” ci è mancato tanto così che avessi un attacco di ansia e buttassi giù tutto	she/her☺ ⬚⬚ || lg(b)t+ ˚｡ ⬚ ⌗ harry ♡ louis ⌗˚ ｡✶˚ 「☮⬚⬚yellow 🫀 (n.) 𝐻𝑎𝑟𝑟𝑦. ⋰. ⬚ 」 mà.05/04/2022 coez.13.03.2022	1	it
it_82	@USER Queer all'inizio era un termine dispregiativo voleva dire checca, frocio ecc...con gli anni a seguire il termine perse il significato originario la comunità lgbt+ se ne appropriò e lo uso per indicare tutte le diversità sessuali.	Nobile punk decaduto, studente di filosofia a tempo perso. Divoratore di libri fumetti e film Ho un quadro che dice cose intelligenti e invecchia al mio posto.	0	it
it_355	@USER Noi froci abbiamo il simbolo arcobaleno perché stiamo sempre a fare l'amore, mica come voi etero che state sempre a farci la guerra	Il Seth Cohen (o forse più Ryan Atwood) di Twitter Tipica faccia da chi vive per l'aperitivo	1	it
it_1640	E comunque non sei lesbica sei pansessuale con preferenza per gli uomini trans quindi 50% queer 40% pan É 20% les… — ma fa 110% 💀 URL	di giorno guardo i video di crime e di notte piango sulla musica indie 🏳️‍🌈 / “bdv” perchè nessuno sa leggere @/btwgiuls (IG)	0	it
it_1020	"i forci devono smetterla di farti domande sul cazz0, se volete scopare o fidanzarvi ditelo direttamente, sicuro vi dirà di sì 🥴" - Sicuramente sì #tellonym URL	august slipped away into a moment in time🌾 (he/him)	1	it

Figure 3: *Sample rows from the dataset.*

of reclaimed language, where slurs regularly co-occur with identity markers and community-positive terms (e.g., orgullo, pride, rainbow). This contrast underscores the central challenge of the task: reappropriative intent cannot be deduced from lexical items alone, but rather from their contextual framing, reinforcing the need for models capable of capturing semantic information.

Figure 4: *Most frequent words that appear in messages labelled as non-reclaimed (Label 0). The relative size of each term reflects its frequency in this subset of the data.*

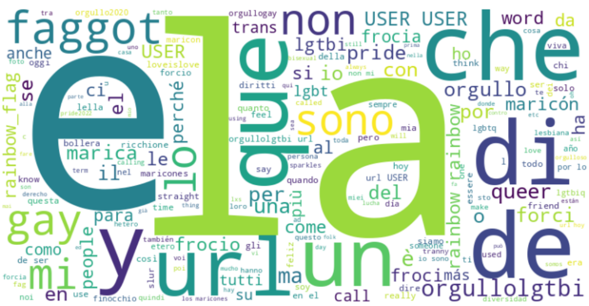

Figure 5: *Most frequent words that appear in messages labelled as reclaimed (Label 1).*

2.3. Description of the Tasks

The MultiPRIDE shared task consists of two binary classification tasks aimed at identifying reappropriative intent in social media messages containing potentially derogatory LGBTQ+-related terms.

- In **Task A**, systems are required to predict whether a message expresses reappropriative intent based on its textual content (i.e., on the text message itself);

- **Task B** extends the task by allowing the use of contextual metadata about the author, specifically the biographical information provided in the *bio* field. This additional context can support the disambiguation between derogatory and reappropriative uses of language.

In this paper, we consider both tasks, and for each task, we provide the relevant details on our approach in the following Methodology section.

3. Methodology

This section illustrates the proposed methodology for addressing the tasks introduced by MultiPRIDE. In particular, it describes the data preprocessing phase, the textual representation techniques applied to the data, and the models selected for classification.

3.1. Text Preprocessing

We applied an identical initial preprocessing pipeline to the English, Italian, and Spanish datasets. This step aimed at normalizing the raw text and ensuring that each message was represented as a single, well-formed string before any task-specific processing.

First, the training data released by the challenge organizers were loaded from CSV files into `pandas` DataFrames using UTF-8 encoding. This choice guarantees a consistent treatment of multilingual content and avoids issues with non-ASCII characters, such as accented letters and symbols frequently present in Italian and Spanish tweets.

We then standardized the textual field corresponding to the message content. Since the original files contained heterogeneous line-break encodings (arising, for example, from different operating systems or from escaped sequences within the exported text), we explicitly normalized all newline patterns. By collapsing all these variants into a single space, we ensure that each message is stored as one continuous line of text, while preserving word boundaries. This is particularly important in our setting, where line breaks may appear both because of platform-dependent formatting and because of how the textual content was exported (e.g., embedded escape sequences in longer strings). Normalizing these artifacts reduces spurious token boundaries and facilitates consistent tokenization across languages.

In addition, we defined a regular expression pattern to identify emoji characters in the text. We used this pattern as a building block for subsequent stages of preprocessing and analysis, where emoji may be either removed, normalized, or explicitly modeled as features depending on the preprocessing strategy. Although language-independent, its early definition allows us to handle emoji in a uniform way in all three corpora. In our experiments, emojis were removed under the intensive preprocessing pipeline adopted for sparse lexical models (BoW/TF–IDF), while they were preserved and normalized under the mild preprocessing used for XLM-RoBERTa.

Overall, this shared preprocessing step produced a clean, line-break–normalized version of the text field for English, Italian, and Spanish, providing a homogeneous basis for the language-specific and task-specific processing described in the following sections.

3.1.1. Task A

For Task A, which relies solely on the textual content of the message, we first constructed a unified multilingual dataset by integrating the three monolingual corpora (i.e., Italian, Spanish, and English). The workflow of Task A is illustrated in Figure 6 and detailed below.

To obtain the integrated dataset, it was necessary to perform preliminary processing steps on the individual corpora in order to ensure their correct merging. For example, the Italian and Spanish datasets already include a `bio` field describing the author's profile, whereas the English dataset does not. To obtain a coherent structure, we explicitly added an empty `bio` column to the English data and aligned the column order across all corpora. The three datasets were then concatenated into a single DataFrame.

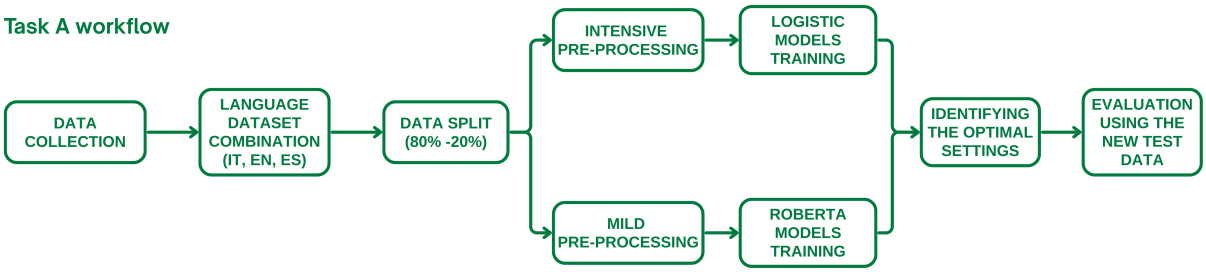

Figure 6: *The workflow for Task A.*

After concatenation, we performed basic sanity checks on the resulting dataset (e.g., verification of the overall shape and column names, and analysis of the label distribution) to ensure that instances from all three languages are correctly merged. This integration step allows us to train a single model over a multilingual corpus.

Starting from our integrated dataset, we applied a unified text preprocessing pipeline to all instances, parametrized by the language of the message. The goal of this pipeline is to normalize surface variation, remove noise typical of social media text, and obtain a lemmatized representation suitable for subsequent modeling.

We defined a set of language-independent cleaning functions, applied in a fixed order to each message. For the intensive preprocessing adopted for sparse lexical models (Bag-of-Words and TF–IDF), we started by removing all emoji characters using a regular expression covering the corresponding Unicode ranges. In contrast, under the mild preprocessing strategy used for XLM-RoBERTa, emoji characters were preserved and later converted into their textual descriptions. Next, hyperlinks, both in `http(s)://` and `www.` format, were removed. We then also removed Twitter-style mentions (strings starting with `@` followed by alphanumeric characters and hyphens), the conventional retweet marker `RT` at the beginning of a message (when present), all numeric sequences (in this task, numbers rarely contribute to detecting reclamatory usage and may interfere with tokenization and lemmatization), and finally all punctuation characters were replaced with whitespace to standardize token boundaries and reduce noise due to idiosyncratic punctuation.

The final step of the preprocessing pipeline consists of language-aware lemmatization combined with stopword removal. For each message, we rely on the `lang` field to select the appropriate language-specific processing rules (English, Italian, or Spanish). Starting from the cleaned text, the message is first segmented into tokens using simple whitespace and punctuation-based heuristics. Each token is then mapped to its corresponding lemma through language-specific morphological rules and lexicons. This procedure reduces inflected forms to a common base form (e.g., reclaimed, reclaiming → reclaim), thereby decreasing sparsity and allowing the model to capture lexical patterns that are invariant to tense, number, or gender.

In parallel, we define a multilingual stopword list by combining three sets of high-frequency function words, one for each language. During lemmatization, any lemma that appears in this multilingual stopword list is discarded. This filtering step removes items such as determiners, conjunctions, and auxiliary verbs, which carry limited semantic content for the reclamation detection task and would otherwise dominate the feature space.

The remaining lemmas are converted to lowercase and concatenated back into a single string, which is stored in a new column of the unified dataset. This lemmatized, stopword-filtered representation constitutes the final input used by our models in Task A and provides a compact, language-consistent basis for learning cross-lingual patterns associated with reclamatory and non-reclamatory uses.

3.1.2. Task B

While Task A relies exclusively on the textual content of the message, Task B additionally exploits contextual metadata about the author, specifically the biographical information provided in the Italian

and Spanish datasets. Because this field is heterogeneous, often noisy, and inconsistently encoded, an additional preprocessing stage was required to clean, normalize, and integrate the biography (`bio`) field with the corresponding message (`text`). The goal was to construct a reliable combined representation that could be used directly as model input. The workflow of Task B is illustrated in Figure 7 and detailed in the following.

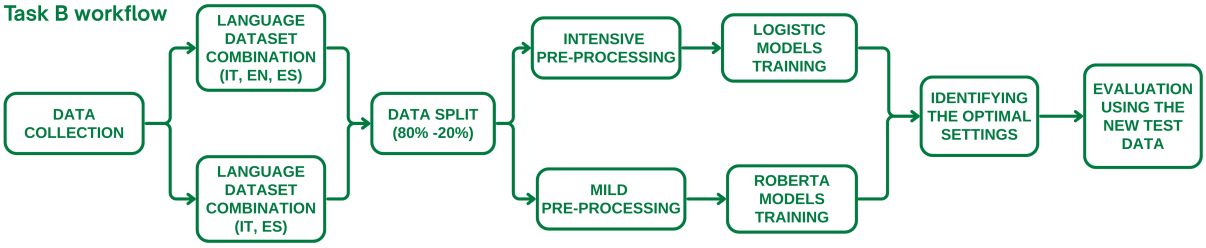

Figure 7: *The workflow for Task B.*

The preprocessing of Task B operates on both the message content and the author biography. Due to inconsistencies in the raw data, we first normalized line breaks and removed problematic encodings that hindered downstream processing. For both the text and bio columns, we applied a uniform sequence of cleaning steps. Since the datasets contain multiple encodings of line breaks, including literal escape sequences, Windows-style line breaks, and true newline characters. All these forms were systematically replaced with a single space. This ensures that both `text` and `bio` fields are represented as uninterrupted one-line strings, similar to the normalization applied in the initial preprocessing. Then, under the intensive preprocessing configuration adopted for sparse lexical models, a broad Unicode pattern was defined to filter out emoji and pictographic symbols from the textual content. Although emojis may signal sentiment or stance, in this setting, their presence was deemed potentially noisy, particularly in the biography field. In contrast, when fine-tuning XLM-RoBERTa, emojis were preserved and converted into their textual descriptions in order to retain potential pragmatic and contextual cues. Finally, both fields were explicitly converted to string type and stripped of superfluous whitespace, reducing variability and preventing alignment issues in concatenation.

The biography field required careful treatment due to the presence of multiple textual placeholders denoting missing content. Therefore, we converted known missing-like patterns into actual `NaN` values, and replaced all resulting `NaN` entries with empty strings. This choice prevents models from mistakenly interpreting placeholder text as meaningful linguistic content once concatenation is performed.

After these operations, the dataset contained no missing values in either field, ensuring that every instance yields a clean concatenated representation.

In order to enforces minimal surface-level uniformity and avoids spurious token boundaries after concatenation, a final cleaning function was applied to both fields, which collapsed multiple whitespace characters into a single space and removed leading and trailing whitespace.

For Task B, the core idea is to provide the model with augmented input that includes both the message content and the author's contextual information. We therefore concatenated the cleaned biography to the cleaned message. Empty biographies simply contribute no additional content, while non-empty biographies enrich the textual signal with relevant contextual cues (such as explicit LGBTQ+ self-identification or references to activist roles). The resulting additional field constitutes the final input representation for Task B.

3.1.3. Validation Set

The validation set provided by the challenge organizers was subjected to the same preprocessing pipeline applied to the training data.

To recap, as for the training portion, all messages were normalized by removing URLs, user mentions, numbers, retweet markers, and punctuation, followed by lowercasing and whitespace reduction. Emoji characters were removed under the intensive preprocessing configuration adopted for sparse lexical

models, whereas they were preserved and normalized under the mild preprocessing used for XLM-RoBERTa. For Task A, each instance was then processed through the same multilingual lemmatization and stopword filtering procedure based on the language specified in the dataset. For Task B, the validation biographies underwent the same cleaning and harmonization steps as the training biographies, including newline normalization, removal of placeholder strings, and conversion of missing values into empty strings before concatenation with the message text.

By mirroring the full preprocessing workflow on the validation set, we ensure that the models are evaluated on data that reflect the exact same transformations applied during training, guaranteeing full methodological consistency.

3.2. Text Representation Models

In order to investigate the most effective textual representations for the tasks under consideration, we explored both sparse and dense representations, which are described below.

3.2.1. Sparse Models

To convert raw multilingual social media text into a format suitable for machine-learning algorithms, we employ embedding representations that capture both semantic and syntactic properties of language. Given the multilingual and socially varied nature of the dataset our embedding strategy prioritizes **cross-lingual comparability**, **sensitivity to lexical variation**, and **robustness to informal language** typical of social media dialogue.

Within this framework, we adopt **sparse lexical representations**, specifically Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF–IDF).

Bag-of-Words provides a straightforward count-based encoding of the textual content, enabling models to capture the distribution of key terms associated with reappropriation without imposing assumptions on word order or syntactic structure. This representation is particularly suitable for social media text, where meaning is often conveyed through distinctive lexical choices, hashtags, or targeted slurs.

Building on this, **TF–IDF** further normalizes word frequencies by penalising terms that are overly common in the corpus and up-weighting more informative ones. This helps distinguish contextually important vocabulary, which is also especially relevant for detecting linguistic reclamation, where the interpretive weight of specific keywords varies across communities and speakers.

Finally, the combination of the BoW and TF-IDF models enables the preservation of a high-resolution view of the lexical surface while simultaneously mitigating noise from tokens that are either extremely frequent or of minimal informative value.

3.2.2. XLM-RoBERTa

In addition to the classical linear models built on Bag-of-Words and TF-IDF, we experimented with a more sophisticated, deep learning approach based on a multilingual Transformer encoder. XLM-RoBERTa (XLM-R) is a state-of-the-art multilingual Transformer trained on 100+ languages using masked language modeling [10]. It provides rich contextual embeddings, allowing the model to capture: cross-lingual semantic patterns, subtle lexical nuances, contextual reappropriation elements (that are difficult for BoW/TF-IDF to encode). The model has been fine-tuned to ensure optimal performance on the binary Reappropriation classification task.

Unlike the sparse lexical representations used in our baseline models (Bag-of-Words and TF-IDF), XLM-RoBERTa requires a different preprocessing strategy. Consequently, we applied a dedicated preprocessing pipeline designed to preserve as much of the original data as possible.

The preprocessing function defined for the XLM-R model performs a minimal set of normalization steps intended to remove non-linguistic noise while leaving intact those elements that may carry contextual information. In particular, hyperlinks in both `http(s)://` and `www.` format are removed, as

they rarely contribute to the target classification task, and instead they frequently introduce irregular token sequences.

Crucially, emoji characters were not removed. Instead, they were converted into their textual description using the `emoji.demojize` function (e.g., `:)` → `smiling face`). This guarantees that the model will continue to have access to important information conveyed through emojis, while still mapping them to a form that is consistently handled by the tokenizer. Since Transformer tokenizers are trained on large multilingual corpora containing emoji descriptions, this strategy allows emojis to contribute meaningfully to the contextual representation learned by the model.

No lemmatization, stopword removal, punctuation removal, or lowercase conversion was performed. Such operations, advantageous for sparse lexical models, would modify the statistics of the subwords on which the transformation models are based and eliminate the morphological information that XLM-R can exploit.

After these first cleaning steps, each message was tokenized using the `xlm-roberta-base` tokenizer from HuggingFace's `AutoTokenizer`. The tokenizer applies Byte-Pair Encoding (BPE), splitting words into multilingual subword units that the pretrained model can interpret consistently across Italian, Spanish, and English. Each processed instance is transformed into the standard input structure required by the model, with sequences truncated or padded to a fixed maximum length to allow efficient batching during training.

This preprocessing pipeline yields a representation that preserves the expressive richness of social media text while ensuring compatibility with the multilingual embedding space encoded by XLM-R. By maintaining contextual signals such as emojis, punctuation, and informal markers, the model is better equipped to identify underlying patterns associated with reclamatory language, which may not be detectable in strongly normalized text.

3.3. Text Classification Models

Following the embedding representations described in the previous section, we evaluate two linear, supervised learning algorithms widely used in text classification: Logistic Regression and Support Vector Machines (SVMs). Both models are well-suited for the high-dimensional and sparse feature spaces produced by the embedding representations introduced in Section 2.2.

Logistic regression is a linear classification model that learns a decision boundary by assigning weights to the input features. In text classification, it offers several advantages: as we said it handles sparse and high-dimensional feature vectors efficiently, it remains relatively interpretable since its coefficients directly correspond to lexical features, and it consistently performs well as a baseline method for linearly separable problems. In particular, in our task, where the goal is to predict a binary Reappropriation label, Logistic Regression represents a reliable and computationally lightweight option.

Support Vector Machines are another standard choice for text classification in sparse feature spaces. The model aims to identify the hyperplane that maximizes the margin between classes, resulting in improved generalization in noisy or imbalanced datasets.

3.4. Training setup

To ensure a consistent and reproducible evaluation framework, all models developed in this project were trained under a unified training setup encompassing data splitting strategies, feature construction, hyperparameter selection, and optimization details. Given the multilingual and multi-task nature of the project, the setup was adapted to the specific characteristics of the two tasks, while maintaining shared methodological principles.

3.4.1. Data Split

For Task A, we rely on the multilingual dataset introduced in Section 1. Following a standard practice in supervised text classification, the corpus was divided into a training set (80%) and a validation set

(20%), using a stratified split to preserve the class distribution of the Reappropriation label. The split is performed with a fixed random seed (equal to 42) to ensure reproducibility across experiments.

We then used a slightly different split for Task B. Each monolingual dataset is partitioned into 60% training and 40% validation. This choice reflects the smaller size of the task-specific corpora and helps to ensure the availability of a sufficiently large validation set for the reliable comparison of models and the tuning of hyperparameters. As in Task A, the division is stratified and deterministic. This validation strategy was adopted to simulate the official evaluation scenario of the shared task, where systems are assessed on an unseen test set provided by the organizers. Using a fixed development split, therefore, ensures methodological consistency and reproducibility across experiments.

3.4.2. Feature Construction and Hyperparameter Choices

The training process builds directly on the embedding representations presented in Section 2.2. For Task A, we employ `CountVectorizer` and `TfidfVectorizer`, configuring them to model not only single words but also short contiguous word sequences of up to three terms. To maintain tractability, the vocabulary is restricted to the 10,000 most frequent features.

For Task B, the training setup builds on the enriched text field produced during preprocessing, which integrates the tweet with its available contextual profile information. To effectively support this input space, we evaluate two TF–IDF configurations: a unigram–bigram model with a maximum vocabulary of 20,000 features, designed to capture short lexical patterns, and a more compact unigram-only model, which often provides greater robustness in low-resource or highly sparse settings.

Hyperparameter tuning is carried out directly on the validation splits. In the multilingual setting, both Linear SVMs and Logistic Regression models are explored using balanced class weights to mitigate the strong label imbalance. Linear SVMs are trained with conservative regularization (`C = 0.001`), while Logistic Regression employs a slightly stronger penalty (`C = 1`) together with an increased iteration limit to ensure convergence (`max_iter = 500`).

3.4.3. Optimization and Validation

All models are implemented using `scikit-learn`, whose deterministic optimization routines are well suited to the sparse, high-dimensional feature spaces produced by our embedding representations. As a consequence, the training process is not organized around epochs or mini-batches. Instead, each optimization step operates on the entire feature matrix, and training terminates once the library's internal convergence criteria are satisfied. Logistic Regression models are optimized with the default solver, whereas Linear SVMs rely on the efficient hinge-loss implementation provided by `LinearSVC`.

Given the size of the multilingual feature matrices and the computational overhead of repeated training, we do not employ k-fold cross-validation. Rather, we adopt a single, fixed development split for all experiments. This split serves multiple purposes: it provides a consistent basis for model selection, supports hyperparameter tuning and threshold calibration, and enables general diagnostic checks (including early detection of overfitting). Throughout this process, performance is systematically assessed using the Macro F1 score, the official evaluation metric of the shared task.

Finally, in order to guarantee reproducibility, all components involving randomness (e.g., data splitting and model initialization) are controlled through a fixed `random_state = 42`. This ensures that results are consistent across runs.

3.4.4. Setup for XLM-R

Unlike the linear models trained on sparse lexical features, the XLM-RoBERTa model required a distinct training setup.

Fine-tuning was performed using HuggingFace's Trainer API with the AdamW optimizer and mini-batch gradient descent. Unlike the scikit-learn models, which operate on the full feature matrix in a single optimization pass, the Transformer is trained for multiple epochs, updating parameters iteratively

over mini-batches. We did not manually set a custom learning rate schedule or warm-up steps. Instead, we used the default values provided by the HuggingFace Trainer.

To address the strong imbalance in the Reappropriation label, we employed a class-weighted loss function. This was implemented through a custom weighted trainer class that injects class weights into the cross-entropy objective during training.

Then, as in Task A, 20% of the multilingual dataset was held out as a validation set using a stratified split. The trainer evaluated the model at the end of each epoch, selecting the checkpoint with the highest Macro F1 score. This ensures comparability with the linear models, which also rely on Macro F1. Training a Transformer on a GPU does not produce exactly the same results every time. However, to make the results as reproducible as possible, the code still sets random seeds wherever the framework allows it.

Once the best hyperparameters were identified (selected empirically using the validation split described above), the model was additionally trained on the full training partition to maximize data usage before generating final predictions for evaluation.

4. Evaluation

This section describes the evaluation protocol adopted in our experiments, with particular attention to the metric required, the Macro F1-score, and to the preliminary validation procedure conducted before the official development labels were released. The goal is to detail how model performance was assessed in a controlled and reproducible manner.

4.1. Macro F1-score

The official evaluation metric for the shared task is the Macro F1-score, defined as the unweighted average of the F1-scores computed for each class. Unlike accuracy, which can be misleading when classes are imbalanced, the Macro F1 gives equal importance to both classes, regardless of their frequency. Why is Macro F1 appropriate for our imbalanced data? Our dataset exhibits a strong imbalance: Label 0 is substantially more frequent, while Label 1 is harder to identify. In such scenarios Macro F1 prevents this by penalizing models that perform poorly on the minority class. Thus, this metric directly reflects the system's ability to detect true instances of reappropriation.

4.2. Preliminary Evaluation on a Training Split

Before the official validation set was released, model development relied on an internal validation split extracted from the training set (80/20). The main purposes of this internal split were to enable early performance estimation, and to guide hyperparameter selection.

The split preserved the original label distribution: the training portion contained 2,390 samples, while the validation portion contained 598 samples, with an identical imbalance of ≈85.7% class 0 vs. ≈14.3% class 1 in both partitions. This confirms that stratification was successful and that the validation subset is representative of the real data distribution.

4.3. Preliminary Results for Task A

4.3.1. Sparse Models

Table 1 reports the Macro F1-scores on the validation split for sparse lexical models using Bag-of-Words (BoW) and TF–IDF representations.

Among sparse models, **Logistic Regression** clearly outperforms LinearSVC, with **Bag-of-Words** yielding the strongest performance.

Given its strong performance, Logistic Regression with Bag-of-Words was further tuned by adjusting the inverse regularization parameter C. The best result was obtained with $C = 5.0$, reaching a Macro F1-score of 0.760.

Table 1
Preliminary results of sparse models for Task A (validation split).

Model	Representation	Macro F1
Logistic Regression	Bag-of-Words	0.758
Logistic Regression	TF–IDF	0.742
Linear SVM	Bag-of-Words	0.699
Linear SVM	TF–IDF	0.691

4.3.2. XLM-R

Finally, the multilingual Transformer XLM-R Base was fine-tuned on the same internal split. Table 2 reports the validation performance of XLM-R across training epochs.

Table 2
Preliminary results of the multilingual Transformer for Task A (validation split).

Epoch	Training Loss	Validation Loss	Macro F1
1	0.679	0.702	0.461
2	0.567	0.513	0.667
3	0.439	0.540	0.740

The model established a peak performance with a Macro F1 score of 0.7403 on the validation set at the conclusion of Epoch 3.

The training exhibited typical behavior for imbalanced classification: the Validation Loss reached its minimum at Epoch 2 (0.5134) and then slightly increased at Epoch 3 (0.5403), even as the target Macro F1 metric improved significantly. This slight divergence, coupled with a consistently dropping Training Loss (from 0.57 to 0.44), suggests an early onset of overfitting or a lack of perfect alignment between the cross-entropy loss function and the Macro F1 metric.

As expected, the contextual multilingual representation provided by XLM-R captures insights more effectively than lexical features.

4.4. Preliminary Results for Task B

Task B uses the same internal split (2,390/598 samples), but additionally allows the use of author metadata for Italian and Spanish.

4.4.1. Sparse Models with Metadata

The integration of metadata leads to consistent improvements over Task A, particularly for Logistic Regression. Table 3 reports the Macro F1-scores of sparse models on Task B.

Table 3
Preliminary results of sparse models for Task B.

Model	Representation	Macro F1
Linear SVM	Bag-of-Words	0.704
Linear SVM	TF–IDF	0.708
Logistic Regression	Bag-of-Words	0.789
Logistic Regression	TF–IDF	0.754

Logistic Regression with **Bag-of-Words** achieves the highest performance among sparse models, confirming the usefulness of metadata for improving minority-class recall.

4.4.2. XLM-R

XLM-R Base was also fine-tuned for Task B. The model achieves the following Macro F1-scores. Table 4 reports the Macro F1-scores obtained by XLM-R on Task B across epochs.

Table 4
Preliminary results of the multilingual Transformer for Task B.

Epoch	Training Loss	Validation Loss	Macro F1
1	0.723	0.722	0.452
2	0.654	0.491	0.734
3	0.482	0.520	0.766

The Macro F1 score improves substantially across every epoch, reaching a peak of 0.7659 in Epoch 3.

4.5. Summary

The internal train–validation split enabled early comparison of models and guided the selection of hyperparameters before training on the full dataset. Overall, the strongest sparse model was Logistic Regression with Bag-of-Words, which achieved a Macro F1 of 0.760 on Task A and 0.789 on Task B. However, the best overall performance across both tasks was obtained by the fine-tuned XLM-R Base model, which consistently outperformed all sparse lexical baselines and achieved the most balanced improvement across both majority and minority classes. Based on these results, the best-performing configurations for each task, the tuned Logistic Regression model, and the fine-tuned XLM-R, will be retrained on the complete training set and subsequently used to generate predictions on the official test data provided by the organizers.

5. Results

This section presents the experimental results obtained on the MultiPRIDE tasks, focusing on a comparative evaluation of the different modeling strategies. We first report a quantitative comparison across models and tasks, and then provide a detailed analysis of the best-performing systems, including qualitative observations and limitations.

5.1. Model Performance Comparison

Table 5 presents the Macro F1-scores obtained by the submitted models on Task A and Task B.

For Task A, which relies exclusively on the textual content of the message, both models achieve competitive performance. Logistic Regression with Bag-of-Words provides a strong baseline, confirming that lexical information alone captures a substantial portion of the signal associated with reappropriative language. However, XLM-RoBERTa achieves higher performance, demonstrating its ability to exploit contextual cues and cross-lingual semantic patterns that are not accessible to sparse representations.

For Task B, which incorporates author metadata through the bio field, performance improves for both models. The Logistic Regression model benefits from the additional contextual information, showing a clear increase in Macro F1-score compared to Task A. Nevertheless, XLM-RoBERTa remains the best-performing model overall, with a larger margin over the sparse baseline. This indicates that the Transformer is particularly effective at jointly modeling message content and author context.

Across both tasks, the results show a consistent trend: while Bag-of-Words combined with Logistic Regression constitutes a robust and efficient baseline, contextualized representations provide a measurable advantage, especially when additional metadata are available.

Table 5
Performance of the submitted models on Task A and Task B (Macro F1 and Accuracy).

Task	Model	Macro F1	Accuracy
Task A	Logistic Regression + BoW	0.758	0.875
	XLM-RoBERTa	0.740	0.895
Task B	Logistic Regression + BoW	0.789	0.893
	XLM-RoBERTa	0.766	0.860

5.2. Best-Performing Models

For both Task A and Task B, the best-performing system in terms of Macro F1 is Logistic Regression with Bag-of-Words.

In Task A, the sparse lexical model achieves slightly higher performance, confirming that surface-level lexical information captures a substantial portion of the signal associated with reappropriative language. However, XLM-RoBERTa remains competitive and demonstrates the ability to model contextual and semantic nuances beyond explicit lexical markers.

In Task B, Logistic Regression again achieves the highest Macro F1 score. Nevertheless, XLM-RoBERTa shows strong performance.

5.2.1. Qualitative Observations

Qualitative inspection indicates that both models perform reliably when reappropriation is explicitly marked through identity-affirming language. Errors tend to occur in highly ambiguous cases or in very short messages, where intent is difficult to infer regardless of the modeling approach.

5.2.2. Limitations

Despite their effectiveness, both submitted models face limitations. Logistic Regression remains constrained by its reliance on surface-level lexical features and limited contextual interaction. XLM-RoBERTa, while more expressive, requires greater computational resources and still struggles with subtle pragmatic phenomena such as irony and sarcasm. Additionally, the absence of author metadata in the English subset restricts the potential gains of Task B for that language.

Overall, the submitted systems illustrate the trade-off between efficiency and expressive power, while confirming the importance of contextual and metadata-aware modeling for the detection of reappropriative language.

The examples reported in Figure 8 highlight typical failure cases of both our models in the detection of reappropriative language. Most errors arise from lexical ambiguity and pragmatic ambiguity. In several cases, the model fails to distinguish between reclaimed in-group usage and genuinely derogatory usage, due to the semantic ambiguity, as the same lexical item can appear either as a slur or in a neutral historical or literal meaning.

5.3. Official Evaluation Results

This subsection reports the final evaluation results computed by the MultiPRIDE organisers on the official labeled test sets. Different from our internal validation experiments, where models were trained on a single multilingual dataset obtained by merging Italian, Spanish, and English data, the official evaluation was conducted using *language-specific training setups*. That is, separate models were trained and evaluated for each language, following the shared task guidelines.

Tables 6 and 7 report the Macro F1-scores obtained by the submitted systems for Task A and Task B, respectively.

id	text	lang
en_1506	If someone calls a straight person a faggot, does it cease to be a slur against queer people?	1
en_802	Upvote for fucking hilariousness and the fact that at least faggot correctly genders me, or downvote for bigotry? I can't make up my mind! Wat do?!	1
en_1789	Weekend is here. What an amazing week this has been. Let's use this extended weekend to celebrate our successes my fellow queer folk.	0
en_264	I think using a term in its original context is fine. * What a queer set of circumstances * What a queer set of circumstances * The young woman had a bounce in her step because she was feeling gay * The brit lit the fag and and took a drag after a long day's work * The young page collected a faggot of brushwood for the fire.	0
en_332	there's video of her calling a gay flight attendant a faggot after he allegedly hit her in the face, even though he wasn't the one who elbowed her, it was another passenger.	0

Figure 8: *Qualitative analysis of misclassified examples.*

Task A. For Task A, which relies exclusively on message text, XLM-RoBERTa consistently outperforms the Logistic Regression baseline in English, Spanish, and Italian. The performance gap is particularly evident in English, where the Transformer-based model shows a clear advantage in capturing contextual and semantic information. In Italian, both models achieve strong results, with XLM-RoBERTa obtaining the highest Macro F1-score overall. These findings confirm that contextualized multilingual representations generalize well even when trained on language-specific data.

Task B. For Task B, which incorporates author metadata, results are reported only for Italian and Spanish, as biographical information is not available for English. In Spanish, the two models achieve comparable Macro F1-scores, with Logistic Regression showing slightly higher recall-driven performance. In Italian, however, XLM-RoBERTa substantially outperforms the sparse baseline, achieving the best overall result across both tasks and languages. This highlights the Transformer's ability to effectively integrate message content with author-level contextual information.

Overall, the official evaluation confirms the trends observed in our internal experiments: sparse lexical models provide strong and stable baselines, while Transformer-based models achieve superior performance, particularly in settings where richer contextual information is available.

Table 6
Official EVALITA results for Task A (Macro F1). Models trained separately for each language.

Language	Model	Macro F1
English	Logistic Regression + BoW	0.529
	XLM-RoBERTa	0.550
Spanish	Logistic Regression + BoW	0.710
	XLM-RoBERTa	0.778
Italian	Logistic Regression + BoW	0.834
	XLM-RoBERTa	0.850

Table 7
Official EVALITA results for Task B (Macro F1). Models trained separately for each language.

Language	Model	Macro F1
Spanish	Logistic Regression + BoW	0.720
	XLM-RoBERTa	0.719
Italian	Logistic Regression + BoW	0.832
	XLM-RoBERTa	0.883

6. Conclusions and Future Work

In this work, we presented the systems developed by the Hate Busters team for the MultiPRIDE shared task at EVALITA 2026, addressing the automatic detection of reappropriative intent in multilingual social media content related to the LGBTQ+ community. We explored both sparse lexical models and contextualized Transformer-based approaches, evaluating their effectiveness across two tasks with increasing levels of contextual information.

The experimental results highlight several key findings. First, Logistic Regression with Bag-of-Words features proved to be a strong and reliable baseline across all settings. Despite its simplicity, this model achieved competitive Macro F1-scores, particularly in Italian and Spanish, confirming that surface lexical cues already encode a substantial portion of the information relevant to reappropriative language.

Second, the fine-tuned XLM-RoBERTa model consistently achieved the best overall performance across both Task A and Task B. The advantage of the Transformer-based approach was especially evident in the official EVALITA evaluation, where models were trained separately on language-specific datasets. In this setting, XLM-RoBERTa demonstrated strong generalization capabilities across English, Spanish, and Italian, confirming its ability to capture contextual, semantic, and pragmatic aspects of reappropriative language that go beyond surface-level lexical patterns.

The inclusion of author metadata in Task B further improved performance for both models. In particular, the official evaluation results show that biographical information plays a substantial role in clarifying reclaimed versus derogatory language, especially in Italian. This confirms that reappropriative intent is often linked to speaker identity and social positioning, which cannot be deducted from message text alone.

Despite these positive results, several limitations remain. From a data perspective, the dataset is strongly imbalanced, with reclaimed instances representing a minority class, which continues to pose challenges for recall. In addition, the absence of the bio field for English limits the potential effectiveness of Task B for that language and reduces cross-lingual comparability. From a methodological point of view, our internal experiments relied on a merged multilingual training setup, while the official evaluation employed language-specific training. While both settings demonstrate consistent trends, this discrepancy indicates the potential value of a more systematic comparison between multilingual and monolingual training approaches.

Future work can explore multiple directions. First, extending the availability of author metadata to English could significantly improve performance in Task B and enrich the model usage. Secondly, the employment of language-specific Transformer models, as opposed to a unified multilingual architecture (as is the case with XLM-R), may prove more effective in capturing the fine-grained linguistic and cultural particularities. Finally, addressing data imbalance through targeted data augmentation or cost-sensitive learning may further improve minority-class detection. In addition, future work may explore the integration of Large Language Models, which have demonstrated strong capabilities in modeling pragmatic inference and contextual reasoning. Their ability to leverage instruction tuning and few-shot learning could further improve the detection of subtle forms of reappropriative language.

Overall, this work confirms the importance of contextual and metadata-aware modeling for the detection of reappropriative language and provides a solid baseline and reference point for future research on intent-sensitive text classification in multilingual and socially grounded settings.

Code and Data Availability

For reproducibility purposes, the source code is publicly available at: https://github.com/camillagentili/multipride-reappropriative-intent-hatebusters/. To obtain access to the data, we invite the reader to refer to the guidelines published in [6].

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 and Grammarly for grammar and spelling checks, paraphrasing, and rewording. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Popa-Wyatt, Reclamation: Taking back control of words, Grazer Philosophische Studien 97 (2020) 159–176.

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

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

[4] M. G. Worthen, Queer identities in the 21st century: Reclamation and stigma, Current Opinion in Psychology 49 (2023) 101512.

[5] 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.

[6] 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.

[7] 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 Workshop Proceedings, 2018.

[8] J. Mata, E. Gualda, A dataset of spanish tweets on people and communities lgbtqi+ during the covid-19 pandemic 2020–2022, Zenodo, 2025. doi:10.5281/zenodo.14878434.

[9] E. Zsisku, A. Zubiaga, H. Dubossarsky, Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination, in: Proceedings of the 16th ACM Web Science Conference (WebSci '24), 2024, pp. 241–249.

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 8440–8451.