Ghavidel-Rajabi at MultiPRIDE: Identity, Toxicity, or Complexity? A Language-Specific Feature Selection Approach to Reclamatory Intent Detection

Houman Rajabi[1,*,†], Fatemeh Ghavidel[1,*,†] and Kourosh Ghahremani[2,†]

[1]*Department of Computer Science, University of Turin, Corso Svizzera 185, 10149 Turin, Italy*
[2]*Faculty of Humanities, Islamic Azad University, Hamedan Branch, Professor Mousivand Blvd, Madani Town, 65181-15743 Hamedan, Iran*

Abstract

Detecting reclamatory intent—the empowering use of potentially offensive terms within marginalized communities—remains a complex challenge in multilingual hate speech detection. This paper presents a Hybrid Fusion architecture developed for the MultiPRIDE @ EVALITA2026 shared task (Task A), predicated on the hypothesis that the linguistic markers of reclamation vary significantly across cultures. Unlike generic multilingual baselines, our approach decouples feature selection for English, Spanish, and Italian. By fusing monolingual BERT embeddings with language-optimized feature vectors, we isolate distinct "drivers" of reclamation: syntactic complexity in Italian, sentiment-toxicity gaps in Spanish, and identity-based pronoun usage in English.

The proposed pipeline utilizes Random Forest to prioritize the most predictive sociolinguistic features, yielding Macro F1-scores of 0.8981 for Italian and 0.6960 for Spanish, but only 0.5460 for English. Our results demonstrate that while explicitly modeling sociolinguistic nuance drives highly competitive performance in Romance languages, the English results reveal a critical "Identity Trap." Our error analysis demonstrates that the model's reliance on first-person pronouns fails to distinguish empowering reclamation from the reporting of victimization. Lacking the semantic depth to resolve this ambiguity, the English model collapses into a hyper-conservative classifier with negligible recall, highlighting that surface-level features are insufficient for English and necessitating semantic role labeling to distinguish agents from targets.

Keywords

MultiPRIDE, Reclamatory Intent Detection, Language-Specific Feature Selection, Hybrid Fusion, LGBTQ+ Discourse

1. Introduction

The automated detection of hate speech has become a cornerstone of content moderation, yet it frequently struggles with the nuances of reclaimed language—the phenomenon where marginalized communities reappropriate pejorative terms for self-empowerment. In LGBTQ+ discourse, terms like "queer" or "faggot" function simultaneously as slurs when used by outsiders and as badges of identity when used by insiders. This duality creates a "reclamation paradox" where the semantic toxicity of a word contradicts the pragmatic intent of the speaker. For standard Natural Language Processing (NLP) models, which often rely on lexical cues and surface-level toxicity scores, these "bad words in good contexts" create a high risk of false positives, silencing the very voices they are meant to protect [1, 2].

While recent advancements in multilingual Transformers (such as XLM-RoBERTa) have improved generalized sentiment analysis, scholarship suggests they often fail to capture culture-specific pragmatics [3]. A "one-size-fits-all" multilingual approach implicitly assumes that the linguistic markers of empowerment are universal. However, research in cross-lingual transfer indicates that models struggle

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

*Corresponding author.
†These authors contributed equally.
houman.rajabi@edu.unito.it (H. Rajabi); fatemeh.ghavidel@edu.unito.it (F. Ghavidel); kourosh.ghahremani@yahoo.com (K. Ghahremani)
https://houmanrajabi.github.io/ (H. Rajabi)
0009-0005-2484-2442 (H. Rajabi); 0009-0003-9427-6592 (F. Ghavidel); 0009-0006-0890-7231 (K. Ghahremani)

with tasks requiring deep sociolinguistic competence, often defaulting to a "Western" or Anglicized view of discourse [4, 5].

In this paper, we present a Hybrid Fusion Architecture developed for the MultiPRIDE @ EVALITA2026 shared task (Task A) [6], designed to resolve this paradox. To overcome the "curse of multilinguality"—where massive models dilute performance on nuanced, low-resource tasks [7]—we propose a Language-Specific Feature Selection pipeline. Our methodology avoids imposing rigid priors; instead, we engineer a broad "toolbox" of 61 sociolinguistic features and utilize a Random Forest ranking algorithm to empirically derive the most predictive drivers for each language.

Our contribution is twofold, bridging the gap between linguistic theory and computational implementation:

- **Linguistic Discovery:** By allowing the feature selection process to operate independently per language, we identified distinct mechanisms of reclamation. Our analysis reveals that English reclamation is primarily identity-driven (reliant on self-referential pronouns), whereas Italian reclamation is complexity-driven (characterized by syntactic density).
- **Computational Architecture:** We implement a dual-stream fusion model that integrates these empirically selected features with monolingual BERT embeddings. By processing a dense, language-optimized feature vector alongside contextual embeddings, our system effectively disentangles toxic slurs from empowering speech.

The proposed architecture delivers competitive classification results, yielding Macro F1-scores of 0.8981 for Italian and 0.6960 for Spanish—a resilient performance for the latter given the higher dialectal variance and data scarcity of the Spanish subset. These results validate the efficacy of a "Shared Toolbox, Different Tools" approach, where the architecture remains consistent, but the sociolinguistic lens adapts to the cultural reality of each language.

Warning: This paper contains examples of explicitly offensive content.

2. Dataset and Preprocessing

2.1. Dataset and Resources

The experimental data was provided by the MultiPRIDE @ EVALITA 2026 shared task organizers. To ensure proper attribution of the aggregated resources, we acknowledge the primary collections from which the monolingual sub-corpora were derived or adapted. The Italian data is primarily drawn from the TWITA collection [8], while the Spanish subset originates from the LGBTQI+ Dataset 2020-2022 [9]. For English, we note that the annotation schema and linguistic definition of reclamation align with the theoretical frameworks proposed by Zsisku et al. [10], focusing on the disambiguation of empowering usage from discrimination.

- **Label 1 (Reclaimed):** The use of potentially offensive terms for self-empowerment or community bonding (e.g., "I am a faggot and proud").
- **Label 0 (Non-Reclamatory):** This category encompasses not only hate speech and toxic attacks but also neutral usage, and instances where the term is used without clear empowering intent.

The dataset was constructed by the organizers using the FATA (First Ask Then Act) annotation protocol, which utilizes community-based annotators to capture the "insider" pragmatic nuance often missed by standard crowd-workers [6].

2.2. Class Imbalance and Training Statistics

A comprehensive analysis of the training data revealed severe class imbalance, confirming that reclamation is a relatively rare phenomenon even within LGBTQ-focused discourse. The English subset proved most challenging, presenting the highest volume of data but the lowest density of positive instances. The specific distribution of the training set used for our feature engineering pipeline is as follows:

- **English:** 1,026 total samples, containing only 88 positive instances (8.5%). This extreme sparsity (approx. 1:11 ratio) poses the most significant challenge for the model.
- **Spanish:** 876 total samples with 133 positive instances (15.1%), representing a moderate imbalance (approx. 1:6.5 ratio).
- **Italian:** 1,086 total samples with 207 positive instances (19.0%), making it the most balanced subset (approx. 1:5.2 ratio), though still heavily skewed toward the non-reclaimed class.

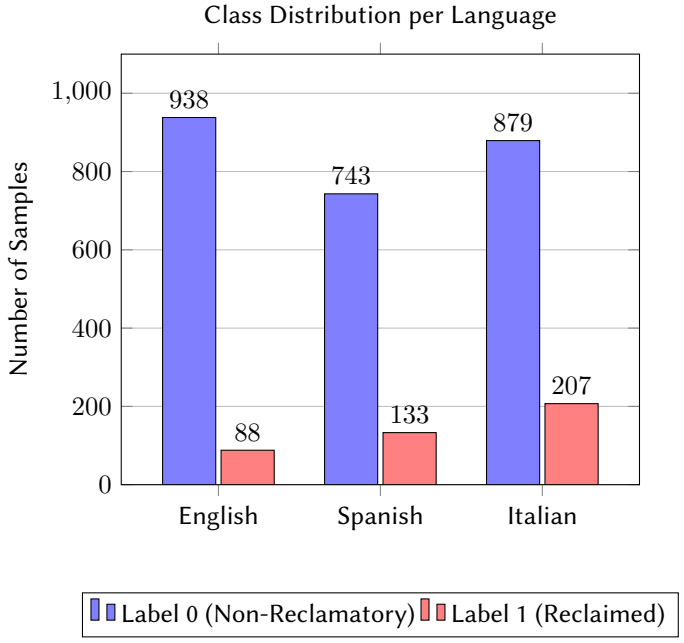

Figure 1: Distribution of Reclamatory (Label 1) vs. Non-Reclamatory (Label 0) instances across languages. The severe sparsity in the English subset (8.5% positive class) highlights the challenge of the task compared to Italian (19.0%).

2.3. Linguistic Feature Extraction

Our methodology prioritizes linguistic retention over aggressive cleaning. Unlike standard pipelines that strip punctuation or "noise," we posit that these elements contain vital signals of tone and intent. Consequently, raw text was fed directly into spaCy Large pipelines (`en_core_web_lg`, `es_core_news_lg`, `it_core_news_lg`) without prior regex filtration for URLs or special characters. This approach allowed us to extract robust syntactic dependencies and Part-of-Speech (POS) tags, generating a feature space that captures structural complexity—a key hypothesis for our Italian model—without losing the "messy" context often found in social media text.

Table 1
Top 5 most predictive features per language identified by Random Forest. Note the prominence of structural features (length/counts) for Italian versus personal/sentiment markers (`first_person_ratio`) for English.

Rank	English Feature (Identity)	Imp.	Spanish Feature (Hybrid)	Imp.	Italian Feature (Complexity)	Imp.
1	`first_person_ratio`	0.0456	`hate_speech_score`	0.0597	`text_length`	0.1139
2	`vader_compound`	0.0384	`hashtag_count`	0.0514	`char_count`	0.1129
3	`caps_ratio`	0.0375	`char_count`	0.0508	`hate_speech_score`	0.1128
4	`self_reference_ratio`	0.0364	`text_length`	0.0506	`word_count`	0.0678
5	`pos_ADP`	0.0321	`lgbtq_term_density`	0.0421	`sentiment_hate_gap`	0.0518

2.4. Feature-Aware Oversampling

To prevent the Random Forest feature selection algorithm from overfitting to the majority class (Non-Reclaimed), we implemented a Targeted Oversampling strategy within the training loop. Rather than simple duplication, we utilized a dynamic sampling approach with a target ratio of 3.0 (reducing the imbalance to 1 minority sample for every 3 majority samples). This specific threshold was chosen to provide the model with sufficient exposure to the minority class patterns (Label 1) without creating an artificial 50/50 balance that ignores the real-world rarity of the phenomenon.

2.5. Task Subjectivity and Theoretical Limits

To contextualize the difficulty of the classification task, it is necessary to acknowledge the inherent subjectivity of defining "reclamation." The use of the FATA protocol [6] by the organizers highlights that ground-truth labels are not objective facts but community-derived consensuses. This subjectivity introduces a "theoretical upper bound" on model performance, particularly evident in the English subset. Qualitative analysis suggests that English reclamation—often reliant on subtle identity markers rather than explicit lexical changes—suffers from higher annotator ambiguity than Italian, where reclamation is frequently marked by distinct structural complexity. This variance in human consensus correlates directly with our model's downstream performance:

- **Italian:** The stability of linguistic markers (complexity/length) aligned with higher model confidence, yielding a robust Macro F1 of 0.8981.
- **English:** The high fluidity of the boundary between "slur" and "reclaimed term" resulted in significant class overlap, explaining the model's struggle to exceed a Macro F1 of 0.5460.

Consequently, the errors in the English subset should be viewed not merely as architectural failures, but as a reflection of the sociolinguistic ambiguity inherent to the dataset itself.

3. Methodology

We propose a Hybrid Fusion Architecture that operates on the premise that sociolinguistic context is not fully captured by standard pre-trained language models. Our system decouples linguistic feature extraction from semantic representation, merging them via a deep fusion mechanism.

3.1. System Architecture

The model utilizes a dual-stream neural network structure processing the input text T through two parallel pathways:

Stream A: Contextual Semantic Embedding The raw text is processed by a monolingual BERT encoder to capture deep semantic dependencies. We utilized distinct checkpoints for each language (e.g., `bert-base-uncased` for English, `dccuchile/bert-base-spanish-wwm-uncased` for Spanish, and `dbmdz/bert-base-italian-uncased` for Italian) to ensure cultural alignment. We extract the pooled output or the mean of the last hidden state, yielding a semantic vector $h_{BERT} \in \mathbb{R}^{768}$.

Stream B: Sociolinguistic Feature Vector Simultaneously, the text undergoes the feature extraction pipeline producing a normalized vector $x_{ling} \in \mathbb{R}^{50}$. This vector is processed by a Feature Encoder MLP designed to project the explicit features into a latent space. The encoder consists of a linear projection to 128 dimensions, followed by Layer Normalization, ReLU activation, and Dropout ($p = 0.5$), yielding a sociolinguistic representation $h_{MLP} \in \mathbb{R}^{128}$.

Deep Fusion Mechanism Unlike simple concatenation, we implement a Deep Fusion Block to model the non-linear interactions between semantic and linguistic signals. The representations are concatenated ($v_{joint} \in \mathbb{R}^{896}$) and passed through a fusion network:

$$h_{fused} = \text{LayerNorm}(\text{ReLU}(W_f \cdot v_{joint})) \tag{1}$$

where the fusion layer projects the dimension down to 256. The final classification is performed via a linear head on this fused vector.

3.2. Linguistic Feature Engineering

We implemented a "Shared Toolbox" approach, extracting a superset of 61 candidate features using **spaCy Large** pipelines and the `twitter-xlm-roberta-base-sentiment` model. The features span three sociolinguistic categories:

- **Morphosyntactic:** Syntactic depth, Part-of-Speech densities, and sentence complexity metrics.
- **Sentiment Dynamics:** The "Sentiment-Hate Gap," serving as a proxy for irony or reclamation.
- **Identity Markers:** Frequencies of self-referential pronouns (1st person) vs. other-referential pronouns.

3.2.1. Derived Sociolinguistic Metrics

To operationalize the "reclamation paradox"—where the semantic toxicity of a term contradicts its pragmatic intent—we engineered composite features that quantify the dissonance between conflicting linguistic signals. Let $S_{raw} \in [-1, 1]$ be the sentiment score derived from XLM-RoBERTa, which we normalize to $S = \frac{S_{raw}+1}{2} \in [0, 1]$. Let $T \in [0, 1]$ be the hate speech probability, R_{self} be the self-reference ratio (1st vs. 2nd/3rd person pronouns), and $E \in \mathbb{N}$ be the count of empowerment keywords. We defined the following heuristic metrics:

1. **Paradox Score (P):** Quantifies the coexistence of positive sentiment and low toxicity, a hallmark of successful reclamation. This feature penalizes toxic positive statements (e.g., sarcasm) by weighting sentiment against the non-toxicity probability.

$$P = S \times (1 - T) \tag{2}$$

2. **Sentiment-Hate Gap (G):** Measures the explicit mathematical distance between the model's perceived sentiment and its toxicity detection. High gaps ($S \approx 1, T \approx 0$) serve as a proxy for the reappropriative shift.

$$G = S - T \tag{3}$$

3. **Reclamatory Intent Score (I_{rec}):** A composite heuristic that weights the Paradox Score (P) with explicit identity (R_{self}) and empowerment markers (E). This scalar value acts as a high-level indicator of "pride" usage.

$$I_{rec} = \frac{P \times (1 + R_{self}) \times (1 + E)}{4} \tag{4}$$

4. **Linguistic Complexity (L_c):** To test the hypothesis that reclamation in Romance languages involves higher syntactic elaboration, we derived a complexity score based on average word length (W_μ) and the unique word ratio (UWR).

$$L_c = W_\mu \times \text{UWR} \tag{5}$$

3.3. Model Complexity and Reproducibility

The total trainable parameter count for the Hybrid Fusion architecture is approximately 110.2 million. The majority of these parameters belong to the fine-tuned BERT backbone (approx. 109M), while the language-specific feature encoders and fusion layers introduce a lightweight overhead of fewer than 300,000 parameters. All models were trained on a single NVIDIA GeForce RTX 5090 GPU (32GB) using a batch size of 32 and the AdamW optimizer.

4. Results

4.1. Performance Analysis: The Explicit-Implicit Divide

Table 2 presents the classification performance of the Hybrid Fusion architecture. The results reveal a sharp divergence in model efficacy, directly correlated with the specific "drivers" of reclamation in each language.

While the model achieved robust performance in Italian (Macro F1: 0.8981) and moderate success in Spanish (Macro F1: 0.6960), it struggled significantly with English (Macro F1: 0.5460). Crucially, the Class 1 (Reclaimed) Recall for English was only 0.10. This indicates that while the model could identify non-reclaimed tweets with high precision, it failed to identify 90% of actual reclamation instances.

This disparity provides empirical evidence for our core hypothesis: Sociolinguistic features (e.g., pronouns, sentiment, sentence length) are strong predictors for explicit/argumentative reclamation (Italian) but fail to capture the implicit/contextual nature of English reclamation.

Table 2
Classification Performance and Primary Drivers

Language	Macro F1	Class 1 (Reclaimed) F1	Class 1 Recall	Primary Feature Driver (Rank #1)
Italian	0.8981	0.8333	0.7914	`text_length` (Complexity)
Spanish	0.6960	0.4638	0.3636	`hate_speech_score` (Toxicity)
English	0.5460	0.1463	0.1017	`first_person_ratio` (Identity)

4.2. Feature Importance and Sociolinguistic Drivers

To understand the mechanics behind these performance differences, we analyzed the Feature Importance rankings generated by the Random Forest selector. This analysis exposes the specific linguistic strategies the model attempted to leverage for each language.

4.2.1. Italian: The "Manifesto" Pattern (Complexity + Low Toxicity)

Italian reclamation was the easiest for the model to detect because it follows a distinct structural pattern.

- **Top Features:** `text_length` (Rank 1) and `char_count` (Rank 2) were the dominant predictors, followed closely by `hate_speech_score` (Rank 3) and `sentiment_hate_gap` (Rank 5).
- **Analysis:** The strong predictive power of the Sentiment-Hate Gap (Rank 5) contradicts the assumption that text length dilutes sentiment signals. Instead, the model learned a specific profile: High Complexity + Low Toxicity = Reclamation.

This suggests that Italian users reclaim slurs through "Manifestos"—long, well-reasoned, non-toxic arguments defending the community. The explicit nature of this style makes it highly detectable by feature-based models.

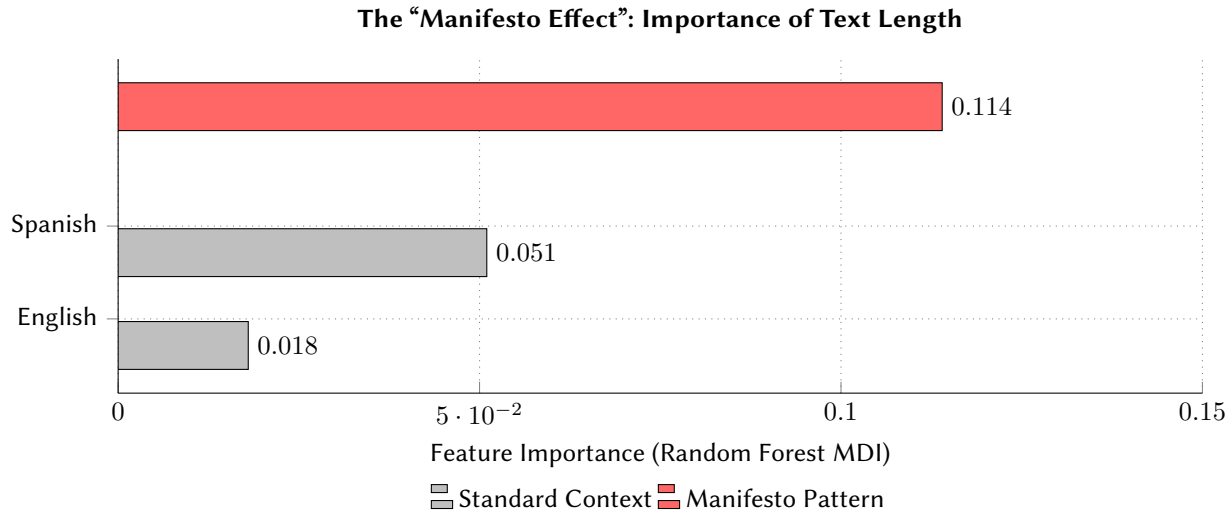

Figure 2: Feature Importance of “Text Length” across Languages. The Italian model (red) relies heavily on length (0.114), validating the “Manifesto Theory” of argumentative reclamation. In contrast, English and Spanish (gray) show significantly lower reliance on length, indicating implicit or toxicity-based drivers.

4.2.2. English: The Anomaly of Feature Poverty

In English, the Random Forest selected `first_person_ratio` as the primary feature (Rank #1), yet this selection coincided with a Recall of only 0.10 for the Reclaimed class. This discrepancy reveals a critical limitation: identity markers are necessary but not sufficient. While explicit reclamation often involves self-reference (e.g., “I am a faggot and proud”), the vast majority of non-reclaimed hate speech reporting also involves self-reference (e.g., “I got punched in the neck and called faggot in junior year of high school”). The model’s reliance on this feature indicates a state of “feature poverty” in English—it lacked the strong structural signals found in Italian (Feature Importance 0.114) and defaulted to weak identity signals (Feature Importance 0.046). Consequently, the model could not confidently distinguish Agents (reclaimers) from Objects (victims), resulting in a high false negative rate where genuine reclamation was lost in the noise of victimization reporting.

- **Top Features:** `first_person_ratio` (Rank 1), `vader_compound` (Rank 2), and `self_reference_ratio` (Rank 4).
- **Analysis:** The model prioritized who is speaking (using pronouns like “I”, “me”, “my”) over what was said. However, the catastrophic Class 1 Recall (0.10) proves that identity markers are a poor proxy for reclamation in English.

English reclamation often occurs through irony, in-group slang, or reappropriation directed at others (e.g., “You guys are queer icons”), which lacks first-person markers. By over-relying on “I am” statements, the model missed the vast majority of implicit reclamation, highlighting the limitation of surface-level sociolinguistic features for English.

4.2.3. Spanish: The Toxicity Threshold

Spanish occupied a middle ground, driven by explicit toxicity signals.

- **Top Features:** `hate_speech_score` (Rank 1) and `lgbtq_term_density`.
- **Analysis:** Unlike Italian (driven by structure) or English (driven by pronouns), Spanish reclamation was best predicted by the absence of hate speech signals in the presence of community terms. The model effectively acted as a “toxicity filter,” identifying reclamation as instances where slurs appeared without the accompanying venom of hate speech.

4.3. Error Analysis

To investigate the limitations of the Hybrid Fusion architecture, we conducted a qualitative analysis of misclassified samples (20 per language). This analysis reveals that the "drivers" identified in Section 4.2—while generally effective—create specific "blind spots" where the selected features conflict with the semantic reality of the text.

Table 3 summarizes the four primary failure modes identified across the languages.

Table 3
Feature Conflict Analysis in Misclassified Samples

Failure Mode	Lang	Example Text (Snippet)	The "Trap" (Feature that Fooled Model)	The "Truth" (Context Model Missed)
Agentivity Confusion	EN	"The one and only time someone called me a faggot... I punched him..." (en_299)	First-Person Pronouns: Model equates "me/I" + "slur" with identity performance.	Semantic Role: User is the Object (victim) of the slur, not the Agent (reclaimer).
Toxicity Barrier	EN	"Proud faggot reporting in..." (en_816)	Lexical Toxicity: The high negative weight of "faggot" in BERT embeddings.	Modifier Ignored: The explicit positive adjective ("Proud") failed to override the slur's toxicity.
Structural Mimicry	IT	"Insulti che vanno da finocchio a checca isterica..." (it_567) *(Insults ranging from faggot to hysterical sissy...)*	Complexity: Long, well-punctuated text (Rank #1 feature for IT).	Genre Confusion: User is reporting hate speech in a formal style, not arguing for reclamation.
Hierarchy Collapse	ES	"Maricones ¿vuestras amigas hetero están compartiendo... #OrgulloLGTBI" (es_423) *(Faggots, are your hetero friends sharing... #LGBTPride)*	Positional Toxicity: Aggressive vocative at the start of the sentence.	Signal Hierarchy: The "softening" hashtag at the end was outweighed by the initial slur.

4.3.1. English: The Semantic Role Failure

The English model's heavy reliance on `first_person_ratio` (Rank #1) creates a critical flaw we term "Agentivity Confusion." The model acts on the heuristic that Slur + Self-Reference = Reclamation.

- **Case Study (en_299):** The theoretical failure of identity markers described in Section 4.2.2 is visibly demonstrated in sample `en_299`. The user writes: "The one and only time someone called me a faggot... I punched him." This is a report of violence and victimization. However, the high density of first-person markers ("me", "I", "we") triggered a False Positive prediction of Reclamation.
- **Implication:** This proves that statistical feature engineering cannot substitute for Semantic Role Labeling (SRL). The model correctly identified the presence of identity markers but failed to distinguish the user as the victim (object) rather than the agent (subject) of the slur.

4.3.2. Spanish: The Signal Hierarchy Conflict

In Spanish, we observed a conflict between the "Contextual Cushioning" features (hashtags) and the "Toxicity" features.

- **Case Study (es_423):** The tweet contains the explicit softening hashtag #OrgulloLGTBI. However, the text opens with an aggressive vocative: "Maricones..." ("Faggots..."). The model classified this as Hate Speech (False Negative).
- **Implication:** This contradicts the assumption that hashtags always act as "safeguards." It suggests a Decision Hierarchy within the model: the raw toxicity of a slur positioned at the start of a

sentence (high saliency) overrides the contextual signal of a hashtag at the end. Conversely, in sarcastic tweets like es_372: "*Feliz día de la mariconeria #LGTBI*" (Happy day of faggotry #LGBT), the model over-trusted the hashtag, leading to a False Positive.

4.3.3. Italian: Genre Confusion

The Italian model's success with `text_length` (F1: 0.8981) led to errors of "Structural Mimicry."

- **Case Study (it_567):** The user provides a detailed list of insults they have received ("from faggot to hysterical sissy"). The text is long, complex, and grammatically standard—matching the "Manifesto" profile the model learned to associate with reclamation.
- **Implication:** The model effectively learned to detect a genre (formal argumentative discourse) but struggles when that same genre is used to report hate speech rather than perform reclamation.

5. Discussion

Our results challenge the conventional reliance in multilingual NLP on uniform representations that implicitly model sociolinguistic phenomena as consistent across languages [3, 7]. While massive models rely on shared semantic representations [7], prior work indicates that this alignment breaks down for high-context pragmatic tasks like hate speech detection [5, 4]. By decoupling feature selection, our error analysis confirms that reclamation is not a universal linguistic act, but a culturally distinct performance that resists a single feature set.

5.1. Italian: Reclamation as Argumentative Discourse

The superior performance of the Italian model (F1: 0.8981) was driven principally by syntactic complexity and text length. We posit that this reflects the specific nature of the Italian dataset, where reclamation frequently manifests as "Counter-Speech."

Unlike casual slang, these instances are characterized by well-structured arguments, correct punctuation, and extended length. The model effectively learned to distinguish "effortful" defense of the community (Reclamation) from "low-effort" slurs (Hate Speech). This supports a "Manifesto Theory" of reclamation in Italian: users reclaiming slurs tend to do so within the context of formal, political argumentation, making the phenomenon structurally distinct and highly detectable.

5.2. English: The Agentivity Trap and Feature Poverty

Our hypothesis that English is "identity-driven" requires nuance. While the model heavily weighted identity markers (pronouns) as the primary feature (Rank #1), the resulting class-specific F1-score (0.15) demonstrates that this strategy acts as a trap. Unlike Italian, where *how* a statement is constructed (syntactic complexity) signals intent, English reclamation often hinges on semantic roles—specifically, distinguishing *who* does *what* to *whom*.

The error analysis reveals that standard sociolinguistic features cannot effectively distinguish the **Agent** ("I am reclaiming this term") from the **Object** ("This term was used against me"). Lacking access to deep semantic role labeling (SRL), the model failed to resolve this "Agentivity Ambiguity." It effectively collapsed into a conservative majority-class classifier, predicting "Non-Reclaimed" in nearly 90% of cases simply because surface-level identity markers were too ambiguous to serve as a reliable decision boundary. This finding suggests that future work in English reclamation detection must move beyond surface features to incorporate dependency-aware role labeling.

5.3. Spanish: Contextual Cushioning and Architectural Conflict

The intermediate performance of the Spanish model (Macro F1: 0.6960) positions it as a true sociolinguistic hybrid, bridging the structural density of Italian and the reliance on specific markers seen in

English. While the Italian model was driven by "Manifesto" style complexity, the feature analysis for Spanish reveals a multifaceted strategy we term *Contextual Cushioning*.

The dominant predictor, `hate_speech_score` (Rank 1), confirms that the model primarily functions as a "Toxicity Filter," assessing whether the semantic weight of the text aligns with hate speech. Crucially, however, the subsequent feature rankings expose a dual-dependency absent in the other languages. The model relies on explicit Community Markers (`hashtag_count` at Rank 2, `lgbtq_term_density` at Rank 5) supported by moderate Structural Complexity (`char_count` at Rank 3). This suggests that Spanish reclamation is identified not strictly by argumentation (as in Italian), but by the presence of "safe" metadata (hashtags) that cushion the toxicity of the slur.

However, this cushioning is subject to a strict Signal Hierarchy; as seen in error analysis (Table 3), the lexical saliency of a sentence-initial slur often overrides the mitigating weight of trailing metadata, leading to False Negatives.

5.4. The Necessity of Language-Specific Drivers

Our findings demonstrate the risk of "zero-shot" or "joint-training" approaches in sociolinguistics. A model trained primarily on English patterns (looking for identity pronouns) would have failed completely on Italian (which looks for length and argumentation).

Spanish served as a hybrid case, where the Paradox Score (the divergence between lexical toxicity and semantic sentiment) proved most effective. This confirms that while the intent (reclamation) is shared, the mechanism differs: Italian uses Structure, English uses Context, and Spanish uses Sentiment/Toxicity contrast.

5.5. Limitations

The primary limitation of this study is the extreme class imbalance in the English dataset (8.5% positive class), which exacerbated the model's bias toward the majority class. Additionally, our "Manifesto Theory" for Italian may be specific to the Twitter/X platform dynamics in Italy during the data collection period and may not generalize to other platforms like TikTok or Instagram.

6. Conclusion

This study presented a Hybrid Fusion Architecture for the detection of reclaimed slurs in multilingual social media data. By integrating pre-trained BERT embeddings with language-specific sociolinguistic feature vectors, we empirically demonstrated that the linguistic realization of reclamation is not universal, but culturally distinct.

Our results validate the "Manifesto Theory" for Italian, where reclamation manifests as high-complexity, argumentative discourse identifiable by structural features (F1: 0.8981). Spanish results (F1: 0.6960) positioned the language as a hybrid case driven by "Contextual Cushioning" (e.g., hashtags), though error analysis revealed a strict Signal Hierarchy where initial lexical toxicity often overrides explicit positive markers. In contrast, the performance gap in English (F1: 0.5460, Recall: 0.1017) exposes the limitations of explicit feature engineering for high-context pragmatics. We found that while English speakers often reclaim slurs through implicit irony and in-group signaling, current models—driven by identity markers like first-person pronouns—fail to distinguish these acts from hate speech reporting.

6.1. Future Directions

The failure to capture implicit reclamation in English suggests that text-only models have reached a performance ceiling. Future work should prioritize two specific technical advancements:

- **Semantic Role Labeling (SRL):** To resolve the "Agentivity Confusion" (Subject vs. Object) identified in our error analysis, explicit modeling of semantic roles is required to distinguish the agent of reclamation from the victim of hate speech.

- **Contextual Metadata:** Integrating user-level bio descriptions and social graph embeddings to provide the necessary disambiguation context when the text itself is implicit (e.g., "You guys are queer icons").

We conclude that effective sociolinguistic modeling requires moving beyond "Universal Multilingualism" toward architectures that respect the specific structural, hierarchical, and pragmatic drivers of each language.

Declaration on Generative AI

(by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the authors used ChatGPT in order to: check grammar and refine the language style. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, 2017, pp. 78–84.

[2] M. Dynel, The reappropriation of slurs: A corpus-based study of "slut" and "bitch" on reddit, Journal of Pragmatics 173 (2021) 162–177.

[3] D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. C. Piqueras, I. Chalkidis, R. Cui, et al., Challenges and strategies in cross-cultural nlp, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 6997–7013.

[4] A. Lauscher, V. Ravishankar, I. Vulic, G. Glavas, From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4491–4503.

[5] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 907–914.

[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] 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.

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

[9] J. Mata, E. Gualda, A dataset of Spanish tweets on people and communities LGBTQI+ during the COVID-19 pandemic 2020-2022, 2022. doi:10.5281/zenodo.14878434, zenodo Dataset.

[10] 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, ACM, 2024, pp. 241–249.