AVAHI at MultiPRIDE: Multilingual Reclaimed Language Detection via Knowledge Graphs and Retrieval-Augmented Generation

Tania Alcántara[1,2,*,†], Omar Garcia-Vazquez[2,†], Jose A. Torres-León[1,†], Marco Cardoso-Moreno[1,2,†], Diana Jimenez[1,2,†] and Luis Moreno-Mendieta[1,2,†]

[1]*AVAHI, Mexico, City, 06600, Mexico*

[2]*Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico, City, 07700, Mexico*

Abstract

This paper presents a hybrid approach to reclaimed language classification in LGBTQ+ tweets for the MultiPride 2026 shared task. Reclaimed language—historically derogatory terms reappropriated by marginalized communities—requires sophisticated contextual understanding due to its usage-dependent meaning. The system integrates three components: a multilingual Knowledge Graph encoding reclaimed terms across Spanish, English, and Italian; a Retrieval-Augmented Generation system using sentence embeddings for case-based reasoning; and Claude 4.5 Haiku accessed via AWS Bedrock for final classification. The system processes both textual content alone (Task A) and text with author biographical information (Task B). Results show substantial cross-linguistic variation: 33.5% positive predictions for Spanish, 20.7% for Italian, and 9.8% for English. Biographical context provided differential benefits, with Italian showing 14.0% relative improvement and Spanish showing minimal change. Average confidence scores remained high (0.857–0.892) across all configurations. The findings demonstrate the effectiveness of combining symbolic knowledge, neural retrieval, and large language models for context-dependent sociolinguistic classification, while highlighting language-specific differences in the value of biographical context for disambiguation. Warning: This paper contains examples of explicitly offensive content.

Keywords

Reclaimed language, LGBTQ+ linguistics, Retrieval-Augmented Generation, Knowledge graphs, Multilingual classification

1. Introduction

Hate speech targeting the LGBTQ+ community in digital environments has become an increasingly relevant phenomenon due to its profound social, cultural, and psychological implications. Online platforms not only amplify the spread of discriminatory messages but also create spaces where language can be used both as a tool of exclusion and as a form of resistance. In this context, expressions historically considered offensive toward sexual and gender diversity do not always function unambiguously: they are often reclaimed by the LGBTQ+ community itself as forms of identity affirmation, irony, or empowerment.

This phenomenon of linguistic reclamation, in which originally pejorative terms are recontextualized and transformed into expressions of pride, belonging, and resistance, presents a substantial challenge

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.

tania.alcantara@avahi.ai,talcantaram2020@cic.ipn.mx (T. Alcántara); ogarciav2023@cic.ipn.mx (O. Garcia-Vazquez); jose.torres@avahi.ai (J. A. Torres-León); marco.cardoso@avahi.ai (M. Cardoso-Moreno); diana.lopez@avahi.ai (D. Jimenez); enrique.moreno@avahi.ai (L. Moreno-Mendieta)

https://www.linkedin.com/in/talcantaram/ (T. Alcántara); https://www.linkedin.com/in/omar-garcia-vazquez-093128219/ (O. Garcia-Vazquez); https://www.linkedin.com/in/josÃĺ-alberto-torres-00a446140/ (J. A. Torres-León); https://www.linkedin.com/in/cardoso1994/ (M. Cardoso-Moreno); https://www.linkedin.com/in/diana-jimenez-aab038201/ (D. Jimenez); https://www.linkedin.com/in/enrique-mm/ (L. Moreno-Mendieta)

0009-0006-4619-9352 (T. Alcántara); 0009-0002-1205-1166 (O. Garcia-Vazquez); 0000-0003-2704-0216 (J. A. Torres-León); 0009-0001-1072-2985 (M. Cardoso-Moreno); 000000023326557X (D. Jimenez); 0009-0007-7198-6849 (L. Moreno-Mendieta)

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

for automatic content moderation systems. The same terms can express hate or, conversely, pride and community, depending on the discursive context, communicative intent, and speaker identity. This semantic ambiguity makes traditional approaches to hate speech detection, based on simplistic binary classifications, insufficient and potentially harmful, as they may censor legitimate expressions of the LGBTQ+ community. In recent years, Natural Language Processing (NLP) has shown notable progress in the automatic detection of hate speech, particularly through models based on transformer architectures such as mBERT [1] and XLM-R [2], which have demonstrated strong performance in multilingual and multicultural settings [3, 4]. However, most of these approaches tend to focus on binary hate classification and often overlook alternative, ironic, or identity-based uses of traditionally pejorative terms, which can lead to bias and censorship of legitimate LGBTQ+ discourse.

At an international level, competitions such as LT-EDI (Language Technology for Equality, Diversity, and Inclusion) have fostered the development of NLP approaches for classifying homophobia and transphobia in social media comments, demonstrating that transformer-based models significantly outperform traditional methods in multilingual tasks [3, 4]. In the Spanish-speaking world, the HOMO-MEX competition organized by IberLEF (Iberian Languages Evaluation Forum) in 2023 provided the first annotated corpus of Mexican Spanish tweets for the detection of LGBTQ+-phobia. By 2024, the challenge expanded into more complex domains such as song lyrics. These efforts have shown that while Spanish-language transformers such as BETO [5] and multilingual models like mBERT [1] can achieve promising results, they also revealed the difficulty of capturing implicit forms of hate as well as reclamation phenomena and dialectal variation typical of Latin American sociocultural contexts.

In this context, MultiPRIDE [6] (Multilingual Automatic Detection of Reclamation of Slurs in the LGBTQ+ Context) emerges as a shared task within EVALITA 2026, specifically focused on the reclamation of offensive terms toward the LGBTQ+ community across different languages and linguistic varieties. Unlike previous tasks focused exclusively on hate speech detection, MultiPRIDE addresses the semantic ambiguity of historically stigmatizing expressions, distinguishing between discriminatory uses and reclaimed uses for identity, humor, or empowerment purposes. It is a multilingual binary classification task involving data in Italian, Spanish, and English, with the goal of determining whether a term related to the LGBTQ+ context in a sentence is being used with reclaiming intent or not.

This task seeks to provide resources and models that enable fairer and more contextualized content moderation, capable of protecting LGBTQ+ individuals from hate speech without silencing their own forms of expression and self-affirmation. MultiPRIDE invites researchers, professionals, and students to explore the linguistic features and challenges associated with reclaimed language within the LGBTQ+ community, considering both textual content (arguments, slurs, derogatory words, self-labeling, rhetorical figures) and contextual information that can be inferred from user profiles, such as their LGBTQ+ identity.

In this paper, we propose the design and implementation of two deep learning models to address the MultiPRIDE task, incorporating preprocessing and representation strategies tailored to Spanish and its regional varieties. [Placeholder for results and analysis]. Our goal is to advance toward automatic moderation systems that not only detect hate speech but also recognize and respect the linguistic reclamation processes that characterize the diversity and vitality of LGBTQ+ discourse online.

2. State of the Art

The problem of linguistic reclamation in the LGBTQ+ context is closely linked to recent developments in automatic hate speech detection research. Although NLP models based on transformers have demonstrated remarkable capacity to identify explicit expressions of homophobia and transphobia, the literature has consistently shown that these systems present substantial limitations when confronting

ironic, community-based, or reclaimed uses of historically offensive terms [7, 8]. In early work on automatic detection of homophobia and transphobia, such as those promoted by LT-EDI [9, 3, 10, 4], the main objective was to distinguish between offensive and non-offensive content across multiple languages. While these efforts consolidated the use of architectures such as mBERT [1] and XLM-R [2] as the de facto standard for hate detection, the annotation and evaluation schemes were primarily based on binary or multiclass classifications that do not explicitly model the semantic ambiguity of stigmatizing terms. In parallel, resources such as HOMO-MEX [11, 12, 13] revealed that, even in Spanish, the performance of models like BETO [5] and mBERT [1] degrades significantly when discourse contains irony, cultural implicatures, or community-based reappropriations, as frequently occurs in creative domains such as song lyrics. These findings suggest that a significant portion of classification error stems not from the model's inability to detect offensive terms, but from its difficulty interpreting their pragmatic function within discourse [14, 15].

This problem has been addressed more explicitly in recent work on queer language and model biases. Andersen et al. [16] demonstrated that the semantic polarity of terms related to the LGBTQ+ community in Mexican Spanish has changed significantly over time, meaning that models trained on historical data tend to overestimate the offensive character of words that have been partially reclaimed. Complementarily, recent studies have shown that language models tend to label discourse produced by queer speakers as harmful more frequently than heteronormative discourse, even when it contains no explicit attacks, revealing a systematic bias against queer dialect and identity-based self-labeling [17, 18, 19]. This phenomenon extends beyond LGBTQ+ contexts, as similar biases have been documented for African American English and other marginalized linguistic varieties [15, 20].

In the Italian context, the ReCLAIM project [21] has directly addressed the phenomenon of slur reappropriation through the construction of resources and models designed to distinguish between discriminatory and identity-based uses of stigmatizing terms, showing that contextual and discursive information is essential for reliable classification. These results reinforce the need to move from purely lexical or sentiment-based hate detection schemes toward models capable of capturing pragmatic relationships, community affiliation, and discursive positioning [22, 23]. The psychological and sociolinguistic dimensions of slur reclamation have been well documented. Galinsky et al. [22] established that group power mediates the willingness to self-label with derogatory terms, and that such self-labeling can increase both individual and collective perceptions of power while attenuating the stigma associated with these terms. From a linguistic perspective, Anderson and Lepore [24] analyzed the semantic mechanisms through which slurs operate, providing theoretical foundations for understanding how the same lexical items can function differently depending on speaker identity and communicative context. In this framework, MultiPRIDE represents a natural evolution of previous LGBTQ+-phobia detection tasks by shifting the focus from mere hate identification toward explicit modeling of the semantic and pragmatic ambiguity associated with linguistic reclamation processes. By formulating the task as a binary classification between reclaimed and non-reclaimed uses of LGBTQ+ terms across multiple languages (Italian, Spanish, and English), MultiPRIDE directly connects with the findings of LT-EDI, HOMO-MEX, and ReCLAIM, and provides an experimental framework for evaluating the extent to which current models can distinguish between symbolic violence and discursive self-affirmation.

From this perspective, the design of the models proposed in this work is grounded in the accumulated evidence that hate speech detection toward the LGBTQ+ community cannot be reduced to the presence of offensive lexicon, but rather requires incorporating signals of context, identity, and pragmatic function [25]. In this way, our proposal seeks to contribute to the development of automatic moderation systems that, in line with MultiPRIDE's objectives, protect LGBTQ+ individuals from linguistic harm without rendering invisible or penalizing their own forms of expression, resistance, and linguistic reappropriation [26].

3. Methodology

A brief description of the provided datasets for the competition is given in this section; furthermore, a detailed description of the method selected to resolve the problem is provided.

3.1. Task Description

The MultiPride 2026 shared task focuses on the classification of reclaimed language within the LGBTQ+ community on social media. The task invites participants to explore linguistic features and contextual elements associated with language reclamation, a phenomenon where marginalized communities reappropriate terms that were originally used as slurs or derogatory labels.

Participants are encouraged to analyze both the textual content of the inputs—including arguments, slurs, derogatory words, self-labeling practices, and figures of speech—and the contextual information that can be inferred from users' profiles when available, such as community membership and political orientation.

The primary objective is to perform binary classification, determining whether a term related to the LGBTQ+ context within a given sentence is used with reclamatory intent or not. The shared task is organized into two main tasks:

3.1.1. Task A - Textual Content

In Task A, participants are provided exclusively with the textual content of messages. The task can be approached in two distinct modalities:

- **Constrained approach:** Participants must utilize only the provided training data. While additional resources such as lexicons are permitted, the use of supplementary training data in the form of tweets or sentences is prohibited.
- **Unconstrained approach:** Participants may leverage additional training data, including external datasets annotated for reclaimed language. Systems developed using this approach must be clearly identified during run submission and thoroughly detailed in the technical report.

Task A comprises three language-specific subtasks. Participants may choose to work on one, two, or all three languages, with multilingual experiments being particularly encouraged:

- **Subtask A1 - Italian:** Classification of reclaimed language in Italian texts.
- **Subtask A2 - Spanish:** Classification of reclaimed language in Spanish texts.
- **Subtask A3 - English:** Classification of reclaimed language in English texts.

3.1.2. Task B - Contextual Content

Task B extends beyond textual analysis by incorporating contextual information derived from authors' profiles, including biographical data when available. This task is organized into two language-specific subtasks:

- **Subtask B1 - Italian:** Classification leveraging both Italian texts and user profile information.
- **Subtask B2 - Spanish:** Classification leveraging both Spanish texts and user profile information.

It should be noted that biographical information is available exclusively for Spanish and Italian datasets. As with Task A, participants may choose to work on one or both languages, and cross-linguistic analysis is strongly encouraged.

3.1.3. Dataset Examples

The following table shows some examples of samples extracted from the three available languages. It should be noted that the English language samples do not have a user or biography, which prevents task 2 from existing for this language.

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

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

3.2. Corpus Description

The MultiPride 2026 shared task provides a unified training set for both Task A (Textual Content) and Task B (Contextual Content). The training data is distributed across three language-specific CSV files, each corresponding to one of the supported languages:

- `train_it.csv` – Italian language dataset
- `train_en.csv` – English language dataset
- `train_es.csv` – Spanish language dataset

3.2.1. Data Format

Each training file follows a standardized format consisting of five fields, as described in Table 2:

The `id` field serves as a unique identifier for each message in the dataset, enabling traceability and reproducibility of experimental results. The `text` field contains the actual message content extracted from social media posts. The `label` field represents the gold standard annotation, indicating whether the text contains terms used with reclamatory intent.

The `bio` field provides contextual information about the message author through their profile biography. This field is particularly relevant for Task B, where contextual information is leveraged for classification. It is important to note that biographical data is not available for all users; consequently, this field may contain missing values. Additionally, biographical information is available exclusively for Italian and Spanish datasets, while the English dataset does not include this feature.

Finally, the `lang` field specifies the primary language of each message, facilitating language-specific processing and enabling multilingual analysis across the three supported languages.

Table 2
Training data format specification.

Field	Description
`id`	Unique identifier for each message
`text`	The textual content of the message
`label`	Binary annotation indicating whether the text contains reclaimed language (reclamatory intent)
`bio`	User profile biography (when available); this field may contain null values for messages where biographical information is unavailable
`lang`	Primary language of the text (it, en, or es)

3.3. Data Loading and Preparation

The training and test datasets were stored in Amazon S3 cloud storage and accessed programmatically using the AWS SDK for Python (Boto3). A dedicated function was implemented to retrieve CSV files from specified S3 buckets and prefixes, handling potential errors during the data retrieval process.

The data loading procedure consisted of two main phases. First, the training data was loaded from three separate language-specific files: `train_es.csv`, `train_en.csv`, and `train_it.csv`, corresponding to Spanish, English, and Italian datasets, respectively. Upon successful retrieval of each file, the number of rows was verified to ensure data integrity. The three training datasets were subsequently concatenated into a unified dataframe, facilitating multilingual model training and cross-linguistic analysis.

Following the consolidation of training data, descriptive statistics were computed, including the total number of training samples and the distribution of labels across the binary classification task. This analysis provided insights into the class balance within the training corpus, which is crucial for understanding potential biases and for implementing appropriate data handling strategies during model development.

In the second phase, the test datasets were retrieved following an analogous procedure. Three language-specific test files—`es_test.csv`, `en_test.csv`, and `it_test.csv`—were loaded from their designated S3 location. These test sets were maintained separately from the training data to enable proper evaluation of the developed models on unseen data, ensuring unbiased performance assessment across all three languages.

3.4. Knowledge Graph Construction

To effectively capture the complex semantic relationships and contextual nuances associated with reclaimed language in the LGBTQ+ community, a Knowledge Graph (KG) was constructed using the NetworkX library for Python. The KG serves as a structured repository of linguistic knowledge, encoding information about reclaimed terms, their usage contexts, and associated semantic attributes across the three target languages.

3.4.1. Graph Structure and Initialization

The Knowledge Graph was implemented as a directed graph, where nodes represent either reclaimed terms or contextual categories, and edges encode relationships between these entities. The graph structure was designed to facilitate efficient querying and reasoning about term usage patterns and contextual appropriateness.

Upon initialization, the KG was populated with a curated set of known reclaimed terms across all three languages. Each term node was annotated with language-specific metadata, including:

- **Language identifier:** Specifying whether the term belongs to Spanish (es), English (en), or Italian (it)

- **Term type:** Classifying the node as a reclaimed term or context category
- **Context attributes:** Providing information about the typical usage context and semantic connotations of the term

3.4.2. Multilingual Term Repository

The KG incorporates language-specific reclaimed terms based on sociolinguistic literature and community usage patterns:

Spanish terms: The Spanish lexicon includes terms such as *marica*, *maricón*, *bollera*, *tortillera*, *joto*, and *jota*. These terms were annotated with their predominant contexts, ranging from positive LGBTQ+ usage to heavily context-dependent interpretations. Regional variations were also considered, particularly for terms like *joto* and *jota*, which are predominantly used in Mexican Spanish contexts.

English terms: The English repository encompasses widely recognized reclaimed terms including *queer*, *dyke*, *fag*, *faggot*, and *trannie*. Each term was classified according to its reclamation status within the LGBTQ+ community, with particular attention to terms that remain controversial or whose appropriateness varies significantly by context and speaker identity.

Italian terms: The Italian lexicon includes terms such as *frocio*, *checca*, *ricchione*, and *finocchio*. Similar to the other languages, these terms were annotated with contextual information reflecting their complex usage patterns within Italian LGBTQ+ communities.

3.4.3. Context Nodes and Semantic Categories

In addition to term nodes, the KG incorporates abstract context nodes representing different usage environments. These include:

- *lgbtq_community*: Indicating usage within community spaces
- *pride_context*: Associated with pride events and celebrations
- *activist_context*: Related to advocacy and activism
- *derogatory_context*: Indicating pejorative usage
- *neutral_context*: Representing ambiguous or neutral usage

These context nodes enable the encoding of relationships between terms and their typical usage environments, facilitating context-aware classification.

3.4.4. Term Extraction and Matching

The KG provides functionality for identifying potential reclaimed terms within input texts. The extraction mechanism operates through pattern matching, comparing normalized (lowercased) text against the terms stored in the graph. Language-specific filtering ensures that only terms corresponding to the declared language of each message are considered as candidates, preventing false matches from cross-linguistic homographs or similar forms.

3.4.5. Contextual Analysis Framework

A key component of the Knowledge Graph system is its contextual analysis capability, which evaluates the semantic environment surrounding identified terms. This analysis leverages both the textual content and available biographical information to assess the likelihood of reclamatory usage.

The contextual analysis framework employs two sets of linguistic indicators:

Positive indicators: Terms and expressions associated with affirmative LGBTQ+ discourse were identified, including explicit community markers (*pride*, *orgullo*, *LGBTQ+*, *comunidad*, *community*), activist terminology (*activism*, *visibility*, *visibilidad*), and values-oriented language (*love*, *amor*, *equality*, *igualdad*, *diversity*, *diversidad*, *respect*, *respeto*). Additionally, symbolic representations such as rainbow flag emojis and heart emojis were included as positive indicators.

Negative indicators: Conversely, terms associated with hostile or discriminatory discourse were catalogued, including explicit hate speech markers (*hate*, *odio*), oppositional language (*against*, *contra*), and specific forms of prejudice (*homophobia*, *homofobia*, *transphobia*, *transfobia*, *discrimination*, *discriminación*).

For each input message, the contextual analysis function computes the frequency of positive and negative indicators in both the message text and the author's biographical information when available. The combined frequency counts are used to derive a sentiment classification (positive, negative, or neutral) and to assess the presence of LGBTQ+-affirming context. This information provides valuable features for downstream classification models.

3.4.6. Term Information Retrieval

The KG also provides a query interface for retrieving detailed information about specific terms. This functionality allows models to access the stored attributes and contextual annotations for any term present in the graph, enabling informed decision-making during the classification process.

The resulting Knowledge Graph comprises nodes representing both linguistic terms and semantic contexts, providing a structured knowledge base that encodes expert understanding of reclaimed language phenomena across the three target languages.

3.5. Knowledge Graph Enhancement from Training Data

While the initial Knowledge Graph was populated with expert-curated reclaimed terms based on sociolinguistic literature, the diversity and evolution of language use within LGBTQ+ communities necessitates data-driven enrichment. To this end, a learning mechanism was implemented to extract additional patterns and contextual information from the labeled training data, thereby augmenting the KG with empirically observed usage patterns.

3.5.1. Training Data Analysis Strategy

The enhancement process focused exclusively on positive examples—that is, messages labeled as containing reclaimed language (label = 1). This selective approach was motivated by the objective of identifying authentic reclamation patterns as manifested in actual community discourse. By analyzing confirmed instances of language reclamation, the system could capture genuine usage contexts, associated linguistic features, and co-occurring elements that characterize reclamatory intent.

The analysis was conducted in a language-stratified manner, processing Spanish, English, and Italian examples independently. This language-specific approach ensured that the extracted patterns and contextual associations remained linguistically and culturally appropriate, respecting the distinct reclamation dynamics that exist across different linguistic communities.

3.5.2. Feature Extraction from Reclaimed Examples

For each reclaimed example in the training data, multiple levels of linguistic analysis were performed to extract salient features:

Hashtag extraction: Social media hashtags serve as important contextual markers, often explicitly signaling thematic content, community affiliation, and communicative intent. A pattern matching mechanism was employed to identify and extract all hashtags present in reclaimed examples. These

hashtags were normalized to lowercase to facilitate consistent matching and comparison. Common hashtags in reclaimed language contexts include pride-related tags (e.g., *#PrideMonth*, *#Pride2022*), community identifiers (e.g., *#LGBTQ*, *#LGTBI*), and specific event or movement markers.

Term identification: For each training example, the previously described term extraction functionality was applied to identify which reclaimed terms from the KG appeared in the text. This process leveraged the language-specific term repositories established during initialization, ensuring accurate identification of relevant lexical items.

Contextual association: When reclaimed terms were identified in a training example, the system created an association between the term and its observed usage context. For each occurrence, the following information was recorded:

- A textual excerpt (limited to 100 characters) providing immediate context around the term usage
- The binary label confirming reclamatory intent
- The set of hashtags co-occurring in the same message

This contextual information was stored in a term-specific repository, creating a corpus of authentic usage examples for each reclaimed term in the KG.

3.6. Vector Database and Retrieval-Augmented Generation System

To leverage the rich contextual information present in the training data, a Retrieval-Augmented Generation (RAG) system was implemented. This system enables the classification model to retrieve and reference similar examples from the training corpus when making predictions on new instances, effectively implementing a case-based reasoning approach informed by empirically observed usage patterns.

3.6.1. Embedding Model Selection

The foundation of the RAG system is a multilingual sentence embedding model, specifically the `paraphrase-multilingual-MiniLM-L12-v2` [1] model from the Sentence Transformers library. This model was selected for its several advantageous properties:

- **Multilingual coverage:** The model supports over 50 languages, including Spanish, English, and Italian, enabling consistent semantic representation across all three target languages of the shared task.
- **Semantic preservation:** The model is trained to generate embeddings that capture semantic similarity rather than mere lexical overlap, allowing the system to identify contextually similar examples even when surface forms differ.
- **Efficiency:** As a distilled variant (MiniLM), the model offers a favorable balance between representation quality and computational efficiency, facilitating real-time retrieval during inference.
- **Cross-lingual alignment:** The model's training on multilingual parallel corpora enables meaningful similarity comparisons across languages, which is particularly valuable for cross-linguistic analysis and transfer learning scenarios.

3.6.2. System Architecture

The RAG system comprises three primary components: an embedding encoder, a vector index for efficient similarity search, and a repository of training examples with their associated labels. The embedding encoder transforms textual inputs into dense vector representations in a high-dimensional

[1] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

semantic space. The vector index, implemented using the FAISS[2] (Facebook AI Similarity Search) library, enables efficient nearest-neighbor search over large collections of embeddings. The example repository maintains the original training instances and their labels, allowing the system to return complete contextual information for retrieved examples.

3.6.3. Vector Index Construction

The construction of the vector database follows a multi-stage pipeline designed to maximize the informational content of embeddings while maintaining computational tractability.

Text preprocessing and augmentation: For each training example, a composite representation was constructed by combining the message text with available biographical information. Specifically, when user biography was available (i.e., for Italian and Spanish examples in Task B), the text was augmented with the biographical content using a structured format: the message text followed by a `[BIO: ...]` marker containing the biography. This augmentation strategy ensures that the resulting embeddings capture both the immediate textual context and the broader author-level context, which is particularly relevant for distinguishing reclamatory from derogatory usage.

For Task A scenarios, where contextual information is excluded, only the message text is used to generate embeddings, ensuring that the retrieval mechanism respects the task constraints.

Batch embedding generation: The composite texts were processed through the embedding model in batches of 32 examples, balancing memory efficiency with processing speed. The model generates fixed-dimensional dense vector representations (embedding dimension determined by the model architecture) that encode semantic content in a continuous space where semantically similar texts are positioned proximally.

The batch processing approach with progress tracking ensures transparency during the potentially time-consuming embedding generation phase, particularly important when processing large training corpora.

FAISS index construction: The generated embeddings were used to construct a FAISS index employing the L2 (Euclidean) distance metric. Specifically, an `IndexFlatL2` index was utilized, which performs exact nearest-neighbor search without approximation. While approximate methods (e.g., `IndexIVFFlat`) could offer computational advantages for very large datasets, the exact search approach was deemed appropriate given the dataset size and the importance of retrieval accuracy for classification performance.

The embeddings were cast to 32-bit floating-point format before index insertion, adhering to FAISS's memory layout requirements while maintaining sufficient numerical precision for similarity computations.

Metadata preservation: Alongside the vector index, the system maintains parallel data structures storing the complete training examples (as dictionary records) and their corresponding labels. This design enables the retrieval mechanism to return not only similar embeddings but also the full contextual information associated with retrieved examples, including text, biography, language, and label information.

3.6.4. Retrieval Mechanism

The retrieval system accepts a query consisting of a text message and optional biographical information, and returns the k most similar training examples based on semantic similarity in the embedding space.

[2] https://ai.meta.com/tools/faiss/

Query processing: Query texts undergo the same preprocessing and augmentation steps applied during index construction. For Task B (contextual content), queries are augmented with biographical information when available, ensuring consistency between query and index representations. For Task A (textual content only), queries consist solely of the message text, excluding biographical context.

The query text is encoded using the same embedding model employed during index construction, generating a query embedding in the same semantic space as the indexed training examples.

Similarity search: The FAISS index performs efficient nearest-neighbor search, identifying the training examples whose embeddings exhibit minimal L2 distance to the query embedding. The search can be configured to retrieve a specified number of neighbors (k), allowing flexible control over the breadth of retrieved context.

Language-specific filtering: To respect language boundaries and prevent cross-linguistic interference, the retrieval mechanism supports optional language filtering. When a target language is specified, the system initially retrieves a larger candidate set (three times the desired k value) and then filters to retain only examples matching the target language. This over-retrieval and filtering strategy ensures that the final result set contains the desired number of same-language examples, which is particularly important for language-specific subtasks.

Similarity score computation: The L2 distances returned by FAISS are transformed into normalized similarity scores using the formula $s = 1/(1 + d)$, where d represents the L2 distance. This transformation maps distances to a $[0, 1]$ interval where higher values indicate greater similarity, providing an intuitive measure of example relevance.

3.6.5. Task-Specific Retrieval Strategies

The RAG system implements distinct retrieval strategies for Task A and Task B, reflecting the different information modalities available in each task:

- **Task A (Textual Content):** Retrieval is based exclusively on message text, with biographical information explicitly excluded from both query and index representations. This ensures that the system learns to identify reclamation based solely on linguistic content.
- **Task B (Contextual Content):** Retrieval incorporates both textual and biographical information, enabling the system to identify similar examples based on both what is said and who says it. This richer representation allows the model to leverage author-level context for disambiguation.

The resulting vector database provides a comprehensive semantic index of the training corpus, enabling efficient retrieval of relevant examples that inform classification decisions through case-based reasoning.

3.7. Large Language Model Integration

The classification system leverages Claude 4.5 Haiku, a state-of-the-art large language model developed by Anthropic, as the primary decision-making component. The model is accessed through Amazon Bedrock, a fully managed service that provides API access to foundation models. Specifically, the model identifier `us.anthropic.claude-3-5-haiku-20241022-v1:0` was utilized, representing the Claude 3.5 Haiku variant optimized for efficient inference while maintaining strong reasoning capabilities.

3.7.1. Classifier Architecture

The classifier implements a hybrid architecture that integrates three complementary knowledge sources: expert-encoded linguistic knowledge (Knowledge Graph), empirical usage patterns (RAG system), and the general linguistic and cultural understanding encoded in Claude's parameters. This multi-source approach enables the system to leverage structured knowledge, case-based reasoning, and large-scale language understanding in a unified framework.

The classifier maintains references to both the Knowledge Graph and RAG system, enabling dynamic context retrieval during inference. Additionally, it establishes a connection to AWS Bedrock Runtime, which handles the low-level details of API communication, request formatting, and response parsing.

3.7.2. AWS Bedrock Integration

AWS Bedrock provides a standardized interface for accessing multiple foundation models, including the Claude family of models. The integration requires proper AWS credentials and IAM permissions, specifically the `bedrock:InvokeModel` permission. The classifier instantiates a Bedrock Runtime client configured for the appropriate AWS region, enabling programmatic invocation of Claude models.

The Bedrock API follows the Messages format specified by Anthropic, requiring structured request bodies that include:

- **API version identifier:** Specifying the Bedrock-specific Anthropic API version (`bedrock-2023-05-31`)
- **Token budget:** Maximum number of tokens Claude may generate in its response (set to 1000 to accommodate structured JSON output with reasoning)
- **Message history:** An array of conversational turns, in this case containing a single user message with the classification prompt

Model invocation returns a response object containing the generated text, which is subsequently parsed to extract classification decisions.

3.7.3. Prompt Engineering Strategy

A critical component of the system is the prompt construction mechanism, which transforms raw input data and retrieved context into a comprehensive instruction for Claude. The prompt engineering strategy employs several techniques to optimize classification performance:

Task specification and definitions: The prompt begins with explicit role assignment, establishing Claude as an expert in LGBTQ+ linguistics and reclaimed language analysis. This priming encourages the model to activate relevant knowledge and adopt an appropriate analytical perspective.

Crucially, the prompt provides precise definitions of key concepts, particularly the distinction between reclaimed language (label 1) and other forms of LGBTQ+-related discourse (label 0). The definition emphasizes that reclamation requires not merely the presence of potentially reclaimed terms, but their use in a positive, affirming, or community-building context. This definitional clarity helps address a common source of classification errors: the conflation of neutral informational content about LGBTQ+ topics with authentic language reclamation.

Contextual principles: The prompt articulates key principles for classification, emphasizing that:

1. Context is paramount—the same lexical item can function as reclaimed or derogatory depending on usage
2. Community membership indicators provide crucial disambiguation signals
3. Author biographical information, when available, offers valuable context
4. Mere mention of LGBTQ+ topics or terminology does not constitute reclamation

These principles encode sociolinguistic insights about language reclamation phenomena, guiding Claude toward contextually appropriate interpretations.

Knowledge Graph context injection: When the Knowledge Graph identifies potential reclaimed terms in the input text, this information is incorporated into the prompt. Specifically, the prompt includes:

- The list of detected terms from the KG's multilingual lexicon
- A binary indicator of whether LGBTQ+-affirming context markers were identified
- Quantified counts of positive and negative sentiment indicators

This structured information provides Claude with explicit signals about the linguistic content and semantic environment of the message, complementing its own pattern recognition capabilities.

Few-shot learning via retrieved examples: The prompt incorporates a few-shot learning component by including the top-3 most similar training examples retrieved by the RAG system. For each retrieved example, the prompt presents:

- The message text (truncated to 150 characters to manage token budget)
- The gold-standard label, explicitly indicating whether the example represents reclaimed or non-reclaimed usage

This few-shot approach enables in-context learning, allowing Claude to observe concrete instances of classification decisions and adapt its reasoning accordingly. The retrieved examples are selected based on semantic similarity to the test instance, ensuring that the provided examples are contextually relevant.

Test instance presentation: The message to be classified is clearly demarcated within the prompt, along with associated metadata:

- The complete message text
- Author biographical information (for Task B only, when available)
- Language identifier (Spanish, English, or Italian)

This structured presentation ensures Claude can clearly distinguish between contextual information and the actual classification target.

Output format specification: To facilitate automated parsing of Claude's responses, the prompt explicitly requests structured JSON output with three fields:

- `label`: Binary classification decision (0 or 1)
- `confidence`: Numerical confidence score on a 0-1 scale
- `reasoning`: Brief textual explanation of the classification decision

The prompt explicitly instructs Claude to respond with JSON only, without additional markdown formatting or commentary, ensuring parseable outputs. The inclusion of reasoning provides interpretability, enabling qualitative analysis of the model's decision-making process.

3.7.4. Classification Process

For each input instance, the classification process follows a multi-stage pipeline:

Context retrieval phase: The system first queries the Knowledge Graph to identify potential reclaimed terms and analyze contextual indicators. Concurrently, the RAG system retrieves semantically similar training examples. These operations execute independently and can be parallelized if needed.

Prompt construction phase: The retrieved contexts are integrated with the input instance to construct a comprehensive prompt following the strategy described above. Language-specific information ensures appropriate linguistic framing.

Model invocation phase: The constructed prompt is formatted into a Bedrock API request and transmitted to Claude 4.5 Haiku. The model processes the prompt and generates a response containing the classification decision.

Response parsing phase: The system extracts the JSON object from Claude's response using pattern matching. A regular expression identifies JSON structures containing the required `label` field, which is then parsed to extract the classification decision, confidence score, and reasoning.

3.7.5. Task-Specific Processing

The classifier supports both Task A (textual content only) and Task B (textual and contextual content) through a configurable parameter. When processing Task A instances, biographical information is excluded from both retrieval (in the RAG system) and prompt construction, ensuring that decisions are based solely on textual content. For Task B, biographical information is fully integrated when available.

3.7.6. Batch Processing

For efficient evaluation on test sets, the classifier implements a batch processing interface that iterates over all instances in a dataset, applies the classification pipeline to each instance, and aggregates results. Progress tracking via the `tqdm` library provides real-time feedback during long-running inference processes.

The batch processor returns a structured dataframe containing classification results, confidence scores, reasoning, and metadata for each instance. This output format facilitates subsequent analysis, error diagnosis, and submission formatting.

3.7.7. Error Handling and Robustness

The system implements comprehensive error handling to maintain operational stability:

- **API errors:** Network failures, timeouts, and service errors are caught and trigger fallback classification strategies
- **Permission errors:** Initialization checks verify Bedrock access permissions and provide diagnostic messages if permissions are insufficient
- **Parsing failures:** Malformed responses are handled gracefully through multiple parsing strategies of increasing leniency
- **Rate limiting:** While not explicitly implemented, the sequential processing approach naturally respects API rate limits

This robust error handling ensures that the system can complete evaluation even when encountering occasional failures, while logging issues for subsequent investigation.

3.8. System Validation on Training Sample

Prior to evaluation on the official test sets, the classification system underwent validation on a held-out sample from the training data. This validation phase served multiple purposes: verifying the operational integrity of the complete pipeline, assessing the system's performance characteristics, and comparing the effectiveness of Task A (textual content only) versus Task B (textual and contextual content) approaches.

3.8.1. Validation Set Construction

A validation subset of 30 instances was randomly sampled from the complete training corpus using stratified random sampling with a fixed random seed (seed = 42) to ensure reproducibility. The relatively small validation set size was deliberately chosen to enable rapid iteration and system debugging while managing computational costs associated with API calls to the Claude model.

The sampling process maintained the natural distribution of labels within the training data, providing a representative subset that reflects the class balance characteristics of the full corpus. The label distribution within the validation set was explicitly recorded to facilitate interpretation of validation results and to ensure awareness of any class imbalance effects.

3.8.2. Dual-Task Evaluation Framework

The validation process evaluated the system under both task configurations to quantify the added value of biographical context:

Task A validation (textual content only): The classifier processed each validation instance using only the message text, explicitly excluding biographical information from both the RAG retrieval process and the prompt construction. This configuration tests the system's ability to identify reclaimed language based purely on linguistic content, representing the most challenging scenario where contextual disambiguation must rely solely on textual cues.

The classification pipeline generated predictions for all 30 validation instances, producing structured outputs containing predicted labels, confidence scores, and reasoning explanations. The predicted labels were then compared against the gold-standard annotations to compute validation accuracy.

Task B validation (textual and contextual content): The same validation instances were re-processed with biographical information included. For instances where user biography was available, this information was incorporated into both the similarity-based retrieval (enabling identification of contextually similar examples) and the prompt (providing Claude with explicit author context).

This configuration represents the full-information scenario where both linguistic content and author identity markers contribute to classification decisions. The comparison between Task A and Task B performance quantifies the informational value of biographical context for reclaimed language identification.

3.8.3. Performance Metrics and Analysis

For each task configuration, validation accuracy was computed as the proportion of correctly classified instances. Additionally, the predicted label distribution was analyzed to identify any systematic biases toward particular classes (e.g., over-prediction of reclamation or non-reclamation).

The merging of predictions with ground-truth labels facilitated instance-level error analysis, enabling identification of challenging cases where the system failed to produce correct classifications. This granular view supports understanding of system strengths and weaknesses.

3.8.4. Qualitative Analysis

Beyond aggregate accuracy metrics, a qualitative analysis examined individual predictions to assess the interpretability and reasoning quality of the system. For a subset of validation instances (typically 3-5 examples), the following information was reviewed:

- **Input text:** The message content being classified (truncated for display purposes)
- **Ground truth label:** The gold-standard annotation from the training data
- **Predicted label:** The system's classification decision
- **Confidence score:** The system's self-assessed certainty (0-1 scale)
- **Reasoning:** Claude's explanation for its classification decision

This qualitative review served multiple analytical purposes. First, it provided insight into the system's decision-making process, revealing which textual and contextual features influenced classifications. Second, it enabled identification of reasoning patterns that distinguish correct from incorrect predictions. Third, it helped diagnose potential failure modes, such as over-reliance on superficial lexical cues or insufficient sensitivity to contextual nuance.

The reasoning text generated by Claude proved particularly valuable for system refinement, as it made explicit the factors considered during classification. For instance, correct predictions often cited specific indicators of community membership, positive framing, or reclamatory intent, while incorrect predictions sometimes revealed misinterpretations of ambiguous language or insufficient weight given to biographical context.

Here are some examples of quantitative analysis predictions where you can see why the label is being assigned.

1. **Text:** "Quit being a faggot and suck that dick" Oh my I lost it ...
   True: 1, **Predicted:** 0
   Confidence: 0.90
   Reasoning: Quote contains a derogatory slur used in a homophobic context, not a positive reclamation by LGBTQ+ ...
2. **Text:** Un lujo colaborar desde @USER con @USER en la semana de Cine Queer en el marco del mes de la diversi...
   True: 0, **Predicted:** 0
   Confidence: 0.85
   Reasoning: While the tweet contains 'Queer', it is used in an informational context about a film event, not as ...
3. **Text:** come quando "ma frocio non mi sembra un insulto" amore mio se io parte della comunità lgbtqia ti dic...
   True: 0, **Predicted:** 0
   Confidence: 0.90
   Reasoning: The tweet discusses the term 'frocio' as an offensive slur, critiquing someone who claims it's not h...

3.8.5. Comparative Performance Assessment

A key objective of the validation phase was quantifying the performance differential between Task A and Task B configurations. The accuracy difference (Task B accuracy minus Task A accuracy) provides a direct measure of how much biographical information contributes to classification performance on this validation sample.

Several scenarios were anticipated:

1. **Positive differential:** Task B outperforming Task A would indicate that biographical context provides disambiguating information that improves classification, validating the inclusion of this feature in the full system.

2. **Zero differential:** Equal performance would suggest that biographical information, at least in this validation sample, does not substantially aid classification, possibly because textual content alone provides sufficient signal.
3. **Negative differential:** Task A outperforming Task B (an unlikely but possible outcome) would suggest that biographical information introduces noise or causes overfitting to spurious correlations.

The observed performance differential, along with analysis of prediction distributions and individual cases, informed final system configuration decisions and provided insights into the relative importance of different information modalities for reclaimed language classification.

3.8.6. Validation Insights and System Refinement

The validation phase provided actionable insights that guided system refinement:

- **Prompt effectiveness:** Analysis of reasoning patterns revealed which prompt components (definitions, principles, retrieved examples, KG context) most strongly influenced correct decisions, informing prompt optimization.
- **RAG system performance:** Examination of retrieved examples for misclassified instances helped assess whether retrieval was surfacing relevant precedents or introducing misleading analogies.
- **Knowledge Graph coverage:** Validation errors sometimes revealed reclaimed terms or usage patterns not captured in the KG, suggesting areas for knowledge base expansion.
- **Confidence calibration:** Comparison of confidence scores with classification accuracy enabled assessment of whether the system's self-reported certainty correlates with actual correctness.

4. Experiments and Results

This section presents the evaluation of the proposed system on the MultiPride 2026 test sets across all subtasks and languages. The experimental design encompasses both Task A (textual content only) and Task B (textual and contextual content), enabling comprehensive assessment of the system's performance and the contribution of biographical information to classification accuracy.

4.1. Experimental Setup

The evaluation was conducted on the official test sets provided by the shared task organizers. Three language-specific test sets were available:

- **Spanish:** 585 test instances for both Task A and Task B
- **English:** 685 test instances for Task A only (biographical information not available)
- **Italian:** 725 test instances for both Task A and Task B

For each test instance, the system generated binary classification predictions (0 for not reclaimed, 1 for reclaimed) along with confidence scores ranging from 0.0 to 1.0. All predictions were obtained using the complete pipeline described in previous sections, incorporating Knowledge Graph analysis, RAG-based example retrieval, and Claude 4.5 Haiku inference via AWS Bedrock.

4.2. Results by Language and Task

4.2.1. Spanish Results

The Spanish test set, comprising 585 instances, was evaluated under both task configurations. Table 3 summarizes the prediction statistics.

Table 3
Prediction statistics for Spanish test set.

Task	Total	Reclaimed (1)	Not Reclaimed (0)	Avg. Confidence
Task A (Text Only)	585	196 (33.5%)	389 (66.5%)	0.861
Task B (Text + Bio)	585	200 (34.2%)	385 (65.8%)	0.860

For Task A, the system classified approximately one-third of the Spanish test instances as containing reclaimed language (33.5%), with the remaining two-thirds classified as not reclaimed. The average confidence score of 0.861 indicates relatively high certainty in the predictions, suggesting that the system could distinguish clear cases of reclamation from non-reclamation in the textual content.

The inclusion of biographical information in Task B resulted in a modest increase in reclaimed predictions, from 196 to 200 instances (a 2.0% increase in the proportion of positive predictions). This shift suggests that biographical context provided additional evidence supporting reclamatory interpretation for a small subset of instances that were ambiguous based on textual content alone. The average confidence remained stable at 0.860, indicating that the additional contextual information did not substantially alter the system's overall certainty levels.

4.2.2. English Results

The English test set, containing 685 instances, was evaluated exclusively under Task A configuration, as biographical information was not available for this language. Table 4 presents the prediction statistics.

Table 4
Prediction statistics for English test set.

Task	Total	Reclaimed (1)	Not Reclaimed (0)	Avg. Confidence
Task A (Text Only)	685	67 (9.8%)	618 (90.2%)	0.892

The English test set exhibited markedly different prediction patterns compared to Spanish. Only 9.8% of instances were classified as containing reclaimed language, resulting in a highly imbalanced prediction distribution heavily skewed toward the negative class. This substantial difference from the Spanish predictions (33.5% positive) raises several interpretive considerations:

1. **Dataset characteristics:** The English test set may contain a genuinely lower proportion of reclaimed language instances, reflecting either sampling procedures or underlying differences in language reclamation prevalence across social media contexts in different languages.
2. **System conservatism:** The system may exhibit greater conservatism in classifying English instances as reclaimed, possibly due to the Knowledge Graph containing fewer English reclaimed terms, or due to differences in the retrieved training examples for English texts.
3. **Linguistic complexity:** English reclaimed language may present greater classification challenges due to more subtle contextual cues or greater ambiguity in term usage, leading the system to default to the negative class in uncertain cases.

Notably, the English predictions exhibited the highest average confidence score (0.892) across all languages and tasks, suggesting that despite the low positive prediction rate, the system expressed strong certainty in its classifications. This high confidence may indicate that many English test instances presented clear negative cases (e.g., neutral discussion of LGBTQ+ topics without reclaimed terminology), or alternatively, may reflect overconfidence in negative predictions.

4.2.3. Italian Results

The Italian test set, comprising 725 instances, was evaluated under both task configurations. Table 5 summarizes the prediction statistics.

Table 5
Prediction statistics for Italian test set.

Task	Total	Reclaimed (1)	Not Reclaimed (0)	Avg. Confidence
Task A (Text Only)	725	150 (20.7%)	575 (79.3%)	0.865
Task B (Text + Bio)	725	171 (23.6%)	554 (76.4%)	0.857

The Italian predictions fell between the Spanish and English extremes, with 20.7% of Task A instances classified as reclaimed. This intermediate proportion suggests either a genuinely moderate prevalence of reclamation in the Italian test set, or system behavior that lies between its conservative English performance and more liberal Spanish performance.

The addition of biographical context in Task B yielded a more substantial effect for Italian than for Spanish. The number of positive predictions increased from 150 to 171 (a 14.0% relative increase, or 2.9 percentage points in absolute terms). This suggests that Italian biographical information provided more disambiguating value than Spanish biographical information, possibly because Italian bios contained stronger community membership signals or because Italian textual content was more ambiguous in isolation.

The average confidence decreased slightly from 0.865 to 0.857 when biographical information was included, suggesting that while biographical context enabled additional positive classifications, it may have introduced some uncertainty or conflicting signals in certain cases.

4.3. Cross-Language Comparison

Figure 1 visualizes the prediction distributions across languages and tasks, facilitating direct comparison of system behavior. The visualization clearly demonstrates the substantial variation in classification patterns across the three target languages and the differential impact of biographical context.

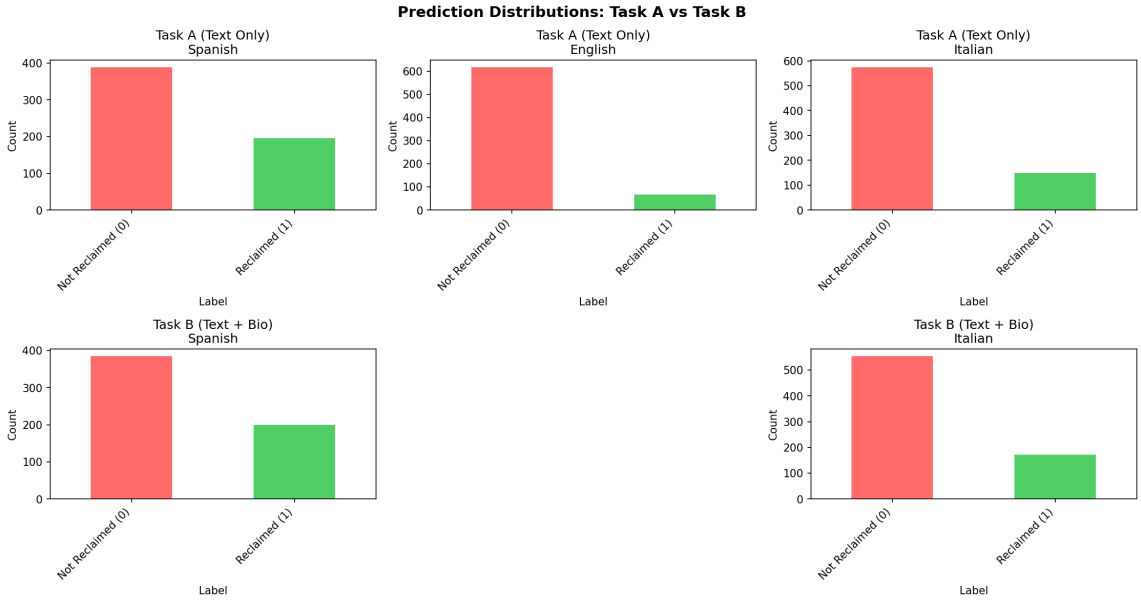

Figure 1: Prediction distributions for Task A and Task B across Spanish, English, and Italian test sets. Red bars represent not-reclaimed predictions (label 0), while green bars represent reclaimed predictions (label 1). The top row shows Task A results for all three languages, while the bottom row shows Task B results for Spanish and Italian (biographical information was not available for English).

4.3.1. Language-Specific Patterns

The visualization in Figure 1 reveals striking differences in prediction patterns across languages:

Spanish: Exhibits a relatively balanced distribution with approximately one-third positive predictions. The consistency between Task A and Task B suggests that textual content alone provides strong classification signals in Spanish, with biographical information playing a supplementary rather than transformative role. The visual similarity between the top-left (Task A) and bottom-left (Task B) panels confirms this minimal differential effect.

English: Shows an extremely imbalanced distribution with minimal positive predictions, as evidenced by the dominant red bar and minimal green bar in the top-middle panel. The stark contrast with Spanish and Italian suggests either genuine dataset differences or language-specific challenges in the classification system. The high confidence scores accompanying this imbalanced distribution warrant further investigation to distinguish between appropriate conservatism and potential systematic bias.

Italian: Demonstrates moderate prediction rates falling between Spanish and English, visible in the intermediate-sized green bar in the top-right panel. The visible increase in positive predictions from Task A (top-right) to Task B (bottom-right) indicates meaningful contribution of biographical context, suggesting that Italian reclaimed language classification benefits substantially from author-level information.

4.3.2. Task A versus Task B

The comparative analysis of Task A and Task B performance, readily apparent in Figure 1, provides insights into the value of biographical context:

Spanish: Minimal visual difference between the top-left and bottom-left panels (0.7 percentage points increase in positive predictions) suggests that Spanish textual content is largely self-sufficient for classification. This may reflect more explicit community markers or clearer contextual cues within Spanish tweets themselves.

Italian: The more noticeable difference between the top-right and bottom-right panels (2.9 percentage points increase) indicates meaningful contribution of biographical information. The visibly larger green bar in the Task B panel demonstrates that Italian reclaimed language usage may be more context-dependent, requiring author identity information to resolve ambiguity.

The modest reduction in confidence when biographical information is included (observed in both Spanish and Italian) may reflect increased complexity in the decision-making process. When biographical context conflicts with or complicates textual signals, the system appropriately expresses lower certainty.

4.4. Confidence Score Analysis

The average confidence scores across all tasks ranged from 0.857 to 0.892, indicating generally high certainty in classifications. Several patterns emerge:

- **English exhibits highest confidence:** The 0.892 average confidence for English, despite (or perhaps because of) the heavily imbalanced predictions visible in Figure 1, suggests either clear negative instances or potential overconfidence in negative classifications.
- **Biographical context reduces confidence:** Both Spanish and Italian show slight confidence reductions when biographical information is included ($0.861 \rightarrow 0.860$ for Spanish; $0.865 \rightarrow 0.857$ for Italian), suggesting that additional context introduces complexity rather than clarity in some cases.
- **Moderate overall uncertainty:** While confidence scores are relatively high, they remain well below 1.0, indicating that the system recognizes the inherent difficulty of the task and expresses appropriate uncertainty rather than overconfident predictions.

5. Discussion

The experimental results reveal several important findings:

Cross-linguistic variation: The substantial differences in prediction rates across languages (9.8% for English, 20.7% for Italian, 33.5% for Spanish), clearly visible in Figure 1, highlight the language-specific nature of reclaimed language phenomena. These differences may reflect actual variation in reclamation prevalence, cultural differences in social media discourse, or differential system performance across languages.

Biographical context value: The comparison between Task A and Task B demonstrates that biographical information provides meaningful but not transformative value. The effect is more pronounced for Italian (14.0% relative increase in positive predictions) than Spanish (2.0% relative increase), as evidenced by the visual differences between panels in Figure 1.

System conservatism: Across all languages, the system tends toward conservative classification, with red bars consistently dominating green bars in the figure, potentially under-predicting reclamation rather than over-predicting it. This conservatism is most evident in English but present to varying degrees across all languages.

High confidence levels: The consistently high confidence scores (>0.85) suggest that the system identifies many clear-cut cases but may benefit from improved calibration to better distinguish varying levels of uncertainty.

These results provide a foundation for understanding system performance and identifying areas for future refinement, including enhanced cross-linguistic knowledge transfer, improved biographical context integration, and better calibration of confidence scores.

Acknowledgments

The authors wish to thank the support of AVAHI.

Declaration on Generative AI

During the preparation of this work, the authors used Writefull's model in order to: Grammar and spelling check. Further, the authors used DeepL Translator in order to: Translate texts from Spanish to English.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[2] 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, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[3] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. P. McCrae, P. Buitelaar, Overview of the shared task on homophobia and transphobia detection in social media comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity, and Inclusion (LT-EDI), 2022, pp. 1–6.

[4] B. R. Chakravarthi, P. K. Kumaresan, R. Priyadharshini, P. Buitelaar, A. Hegde, H. Shashirekha, S. Rajiakodi, M. Á. García, S. M. Jiménez-Zafra, J. García-Díaz, R. Valencia-García, K. K. Ponnusamy, P. Shetty, D. García-Baena, Overview of third shared task on homophobia and transphobia detection in social media comments, in: Proceedings of the Fourth Workshop on LT-EDI (EACL), 2024, pp. 1–10.

[5] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[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] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM), 2017, pp. 512–515.

[8] J. Kurrek, H. M. Saleem, D. Ruths, Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, 2020, pp. 138–149.

[9] B. R. Chakravarthi, Detection of homophobia and transphobia in YouTube comments, International Journal of Data Science and Analytics 18 (2024) 49–68. doi:10.1007/s41060-023-00400-0.

[10] B. R. Chakravarthi, R. Ponnusamy, S. Malliga, P. Buitelaar, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, N. Jindal, Overview of second shared task on homophobia and transphobia detection in social media comments, in: Proceedings of the Third Workshop on LT-EDI, 2023, pp. 1–10.

[11] J. Vásquez, S. Andersen, G. Bel-Enguix, H. Gómez-Adorno, S.-L. Ojeda-Trueba, HOMO-MEX: A Mexican Spanish annotated corpus for LGBT+phobia detection on Twitter, in: Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, 2023, pp. 202–214.

[12] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, S.-L. Ojeda-Trueba, Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish speaking lgbtq+ population, Procesamiento del Lenguaje Natural 71 (2023) 361–370.

[13] H. Gómez-Adorno, G. Bel-Enguix, G. Sierra, J. Vásquez, S. T. Andersen, S.-L. Ojeda-Trueba, Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population, Procesamiento del Lenguaje Natural 73 (2024) 393–405.

[14] T. Davidson, D. Bhattacharya, I. Weber, Racial bias in hate speech and abusive language detection datasets, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 25–35. doi:10.18653/v1/W19-3504.

[15] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668–1678.

[16] S. T. Andersen, S.-L. Ojeda-Trueba, J. Vásquez, G. Bel-Enguix, The mexican gayze: A computational analysis of the attitudes towards the lgbt+ population in mexico on social media across a decade, in: Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, 2024, pp. 178–200.

[17] A. M. Davani, D. Kiela, M. Lambert, B. Xiang, B. Vidgen, T. Thrush, Z. Waseem, Harmful speech detection by language models exhibits gender-queer dialect bias, in: Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, ACM, 2024, pp. 1–13.

[18] T. D. Oliva, D. M. Antonialli, A. Gomes, Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online, Sexuality & Culture 25 (2021) 700–732. doi:10.1007/s12119-020-09790-w.

[19] L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and mitigating unintended bias in text classification, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2018, pp. 67–73.

[20] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5454–5476.

[21] L. Draetta, S. Frenda, A. T. Cignarella, P. Viviana, ReCLAIM It! Exploring Italian slurs reappropriation with LLMs, in: Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, 2024, pp. 1–7.

[22] A. D. Galinsky, K. Hugenberg, C. Groom, G. V. Bodenhausen, The reappropriation of stigmatizing labels: Implications for social identity, in: J. T. Polzer (Ed.), Research on Managing Groups and Teams: Identity Issues in Groups, volume 5, Emerald Group Publishing Limited, 2003, pp. 221–256.

[23] R. Brontsema, A queer revolution: Reconceptualizing the debate over linguistic reclamation, Colorado Research in Linguistics 17 (2004) 1–17.

[24] L. Anderson, E. Lepore, Slurring words, Noûs 47 (2013) 25–48. doi:10.1111/j.1468-0068.2010.00820.x.

[25] F. Elsafoury, Thesis distillation: Investigating the impact of bias in NLP models on hate speech detection, in: Proceedings of the Big Picture Workshop, Association for Computational Linguistics, Singapore, 2023, pp. 53–65. URL: https://aclanthology.org/2023.bigpicture-1.5/. doi:10.18653/v1/2023.bigpicture-1.5.

[26] O. L. Haimson, D. Delmonaco, P. Nie, A. Wegner, Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–35. doi:10.1145/3479610.