HomoCIC at HOMO-LAT 2025: Retrieval Augmented Classification using Transformer Architectures and Vector Knowledgebases

Marco Cardoso-Moreno[1,*,†], Luis Moreno-Mendieta[1,†], Diana Jiménez[1,†], José Alberto Torres-León[1,†] and Jose Valdez-Rodríguez[1,†]

[1]*Instituto Politécnico Nacional, Center for Computing Research, Computational Cognitive Science Laboratory, Mexico, City, 07700, Mexico*

Abstract

This paper presents HomoCIC's approach to the HOMO-LAT 2025 task, focusing on polarity detection in LGBT+ related content from Latin American Reddit posts. Our methodology combines retrieval augmented classification using transformer architectures with vector knowledge databases to address the challenges of cross-dialect hate speech detection. We implemented a comprehensive preprocessing pipeline to handle social media text artifacts, including URL removal, markdown cleaning, emoji conversion to Spanish, and whitespace normalization. Our approach utilizes two embedding models: Amazon's Titan Text Embeddings V2 and a fine-tuned BETO model; the resulting embeddings were stored in an OpenSearch vector database. Classification was performed using k-Nearest Neighbors (kNN) with additional filtering by country and keyword to account for regional linguistic variations across Latin American countries. The fine-tuned BETO model achieved an F1-score of 0.4661, substantially outperforming the general-purpose Titan model (F1-score: 0.3792) by approximately 0.1 points. Our results demonstrate the importance of domain-specific fine-tuning and the feasibility of using retrieval augmented classification for polarity detection in multilingual, cross-dialectal scenarios; however, the moderate performance levels suggest that working with limited training data and cross-dialect classification presents substantial difficulties for current NLP approaches.

Keywords

hate speech detection, transformer architectures, vector databases, BETO, Spanish NLP, retrieval augmented classification, embedding models

1. Introduction

Recent years have seen a rise in hate speech expressions, primarily driven by increased social media activity [1]. The European Union defines hate speech as: "All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic" [2]. Given that hate expressions contain offensive and harmful content targeting specific communities, they tend to cause harm and conflict, and such expressions can easily spread widely due to existing prejudices [3]. Homophobic hate speech holds particular significance, considering that LGBT+ individuals face substance abuse disorders, mental health challenges, workplace discrimination, and restricted access to healthcare services [4, 5]. This becomes even more critical within Mexican society, where drug consumption represents a social problem affecting not only the LGBT+ community [6].

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

†These authors contributed equally.

mcardosom2021@cic.ipn.mx (M. Cardoso-Moreno); lmorenom2021@cic.ipn.mxLuisMoreno-Mendieta (L. Moreno-Mendieta); dianaljl.99@gmail.com (D. Jiménez); jtorresl2019@cic.ipn.mx (J. A. Torres-León); jvaldezr2018@cic.ipn.mx (J. Valdez-Rodríguez)

https://cardoso1994.github.io/ (M. Cardoso-Moreno); https://github.com/JAlbertoTorres (J. A. Torres-León); https://github.com/EduardoValdezRdz (J. Valdez-Rodríguez)

0009-0001-1072-2985 (M. Cardoso-Moreno); 0009-0007-7198-6849 (L. Moreno-Mendieta); 000000023326557X (D. Jiménez); 0000-0003-2704-0216 (J. A. Torres-León); 0000-0002-4572-5713 (J. Valdez-Rodríguez)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-12-10

CEUR-WS.org/Vol-4098/HOMOLAT2025_paper5.pdf

Within this framework, the HOMO-MEX task [7, 8] emerged as part of IberLEF (Iberian Languages Evaluation Forum) [9]. The task's primary goal involves developing Natural Language Processing (NLP) systems capable of identifying LGBT+ related hate speech in Spanish tweets, regardless of how subtle the expression might be. The 2024 edition of HOMO-MEX [8] featured three tracks: a multi-class hate speech detection track mapping tweets to three categories (LGBT+phobic, not LGBT+phobic, and not LGBT+related); a multilabel hate speech detection track with possible classes including Lesbophobia, Gayphobia, Biphobia, Transphobia, Other LGBT+phobia, and Not LGBT+related. The third track focused on classifying song lyrics containing LGBT+phobic hate speech, with classes defined as LGBT+phobic and Not LGBT+phobic. This marked the first inclusion of such a track in the HOMO-MEX task, acknowledging the difficulty of identifying hate speech in songs since detection depends on the context and culture in which the songs were created.

For the year 2025, the HOMO-MEX task has evolved into HOMO-LAT [10], as part of the IberLEF 25 edition [11]; the task was expanded to not only include mexican texts, but instead texts from several countries are included, for instance: Argentina, Chile, Colombia, and Mexico. There are two tracks for this task, in the first one, given a text, the country it came from and a keyword (a hint to what segment of the LGBT+ community the text is directed to), one must determine if the text had either positive (POS), neutral (NEU) or negative (NEG) polarity against the LGBT+ community. Both datasets, train and testing, were conformed by the same countries. Track 2, however, required to perform the same classification, with the difference that in training there were texts from Argentina, Chile, Colombia, and Mexico, while the testing had texts from Bolivia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Uruguay, and Venezuela. Track 1 is considered the Multi-dialect polarity detection track (Multi-class), while track 2 is a Cross-dialect polarity detection (Multi-labeled) task.

This paper presents our approach to tackle these hate expressions against the kLGBT+ community, consisting on the creation of vector databases for later classification of test texts using the k-Nearest Neighbors (kNN) algorithm based on similarity measures between contextual embeddings. The emebeddings were created with two approaches: i) using a propietary model from Amazon with multilingual support and ii) by first fine-tuning a BETO model to this particular dataset and, once trained, get the contextual embeddings from BETO.

The manuscript follows this structure: Section 2 presents a brief literature review for hate speech detection, first providing a general overview and then focusing specifically on the HOMO-MEX task; Section 3 details our approach, covering preprocessing, models and metrics; Section 4 presents the obtained results; finally, Section 5 emphasizes the significance of our approach and indicates future research directions.

2. Literature Review

This section provides a concise overview of existing research concerning hate speech and homophobic language. The review begins with general approaches to the task and subsequently addresses works specifically related to the HOMO-MEX task.

2.1. General Overview

Conventional Machine Learning (ML) models have been frequently combined with Natural Language Processing (NLP) preprocessing strategies. For example, in [12], a voting-based ensemble composed of Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) was utilized in conjunction with character-level and word-level n-grams, as well as syntactic n-grams. This system was designed for the Hate Speech Spreaders (HSSs) profiling task at PAN CLEF 2021 and reported accuracy scores of 0.73 for English and 0.83 for Spanish. Likewise, [13] implemented three tree-based classifiers—RF, Light Gradient Boosting Machine (LightGBM), and CatBoost—optimized using Bayesian search, employing unigram and bigram features, achieving accuracies ranging from 0.85 to 0.87 depending on the model.

Convolutional Neural Networks (ConvNets) have also seen widespread application in hate speech detection tasks. Ribeiro and da Silva [14] introduced a ConvNet architecture for hate speech classification within the SemEval-2019 Task 5. Their model incorporated pre-trained word embeddings such as GloVe and FastText (300 dimensions), yielding F1-scores between 0.48 and 0.69. Similarly, Siino et al. [15] addressed the HSSs task using a ConvNet with a single convolutional layer, reaching an accuracy of 0.79 in a multilingual context (English and Spanish), with language-specific scores of 0.85 and 0.73 for Spanish and English, respectively.

The study in [2] proposed the A-stacking classifier, an ensemble method incorporating a Recurrent Neural Network (RNN) for word embedding generation, a Long Short-Term Memory (LSTM) unit, and a softmax output layer. This architecture was evaluated across multiple datasets under both in-domain and cross-domain configurations. In a related effort, Corazza et al. [16] developed a modular neural model consisting of an RNN layer, a 100-neuron dense layer, and a final output unit. This system supported both word-level and tweet-level representations and was validated on English, German, and Italian datasets.

In the SemEval-2019 competition, the NULI team [17] fine-tuned the Bidirectional Encoder Representations from Transformers (BERT) model to identify hate speech. Their minimal preprocessing pipeline included emoji normalization, hashtag segmentation, and lowercasing, which contributed to their first-place ranking in the competition.

Finally, Caselli et al. [18] introduced HateBERT, a variant of BERT retrained on the Reddit Abusive Language English (RAL-E) corpus for English abusive language detection. Evaluated alongside the original BERT on several datasets, HateBERT consistently outperformed the baseline model, demonstrating the advantages of task-specific pretraining.

2.2. HOMO-MEX Literature Review

The HOMO-MEX task was introduced in 2023 during the IberLEF forum [7]. It involved the construction of a Mexican Spanish corpus of tweets containing terms associated with the LGBT+ community for the purpose of hate speech detection. The vocabulary selection process included slurs, slang, and general terminology gathered from social media platforms such as Twitter, Facebook, and Instagram. Variants of key terms were also accounted for; for example, the word *puto* included derived forms such as *pute* and *putx* (feminized), *putito, putín* (diminutive), and *putote, putón* (augmentative), among others.

Following the compilation of terms, a large-scale web scraping process was conducted, yielding 706,886 tweets in Mexican Spanish. From this dataset, 11,000 tweets were manually annotated into three categories: LGBT+phobic, non-LGBT+phobic, and not relevant to the LGBT+ context. Moreover, under a multilabel classification setting, tweets could be tagged with one or more of the following: Gayphobia, Lesbophobia, Biphobia, Transphobia, or other forms of LGBT+phobia.

Among the submissions to this initial edition, the work of [19] utilized standard NLP preprocessing and feature extraction methods, including Bag of Words (BoW), term frequency (TF), and inverse document frequency (IDF), resulting in TF-IDF representations. The models applied were a Linear Support Vector Machine (LSVM) and a Bagging ensemble using LSVM as the base classifier. Rivadeneira et al. [20] focused exclusively on the multilabel task, training separate classifiers for each LGBT+phobic category. They employed n-gram features from word tokens and weighted TF-IDF BoW representations, using Random Forest and SVM classifiers.

In [21], transformer-based models such as BERT and RoBERTa demonstrated strong performance in both multiclass and multilabel subtasks, effectively identifying LGBT+phobic content. Similarly, the study in [6] employed BERT models for the first track, reporting a Macro F1 score of 0.73. Additionally, Rosauro and Cuadros [22] applied BETO [23], RoBERTuito [24], and mDeBERTa [25]—all BERT-derived models—across both tracks, achieving Macro F1 scores of 0.84 and 0.68, respectively.

3. Proposal

This section provides insight on our proposal for both tracks of the 2025 edition of HOMO-LAT [10].

3.1. Preprocessing

The data for this task consist on Reddit posts from different Latin American countries, including: Argentina, Chile, Colombia, and Mexico for task 1.

Several issues arise when working with this kind of structured data, for instance, the length of posts is variable between each other; since Reddit is a social network there is not a homogeneous, nor formal, writing style, i.e., each user may write as they wish; slang is primarily used; scrapped text tends to come with some formatting tags and marks, links to other resources (news, webpages, images, links to other users or subreddits); etc. Therefore, preprocessing becomes an imperative step that needs to be handled with care.

Our preprocessing pipeline consisted of the following steps (each explained in its own section):

- Remove URLs
- Remove quote markers
- Remove Markdown
- Remove List Prefixes
- Remove Separator Lines
- Remove Sarcasm Tags

Remove URLs Two cases were handled when dealing with URLS: i) plain URLs, for which the URL was simply replaced by the word ENLACE: `https://example.com -> ENLACE`; ii) URLs formatted with markdown, where the user can provide a caption that will be displayed on top of the URL, in this case we decided to remove the URL from the metadata and keep only the captions: `[Caption](url) -> CAPTION`.

Remove quote markers Quote blocks are typically used in Reddit when referring to another post or a given source, a quote block uses (in plain text) the *greater than* (>) symbol as the start of each of its lines, therefore, this removal implicates the detection of lines starting with that symbol and the corresponding removal of it along with any blank characters. This way, only the actual text was preserved.

Remove Markdown Remove all Markdown markers from text.

- Headers: They use the *hash* (#) symbol. There might be one or several hashes, depending on the level of the header
- Text formatting: markers for bold, italics, underlined, strikethrough, superscripts
- HTML entities
- Reddit artifacts and special characters (null or BOM, for instance)
- List Prefixes
- Separator Lines
- Sarcasm tags
- Anonymize Reddit Reference
- Convert Emojis to Spanish
- Limit Repeated Characters
- Normalize Whitespaces

Remove List Prefixes Removal of Markdown list prefixes (-, *, +, 1., etc.) from each line of text.

Remove Separator Lines Some users use their own kind of separator lines to separate sections inside their posts, these vary from person to person but a common example might be: lines with only *equals* (=) or *dash* (−) symbols. Since these separators are user specific there is no automated way of detecting them, our proposal considers a line to be a separator if it contains only one of the following symbols: `-_*=#\+~|/()[]{}`, or if the count of these special characters is larger than the count of alphanumeric characters.

Remove Sarcasm Tags People in Reddit tend to use the `\s` tag to emphazise that the content they are about to express is to be understood as sarcasm; given the nature of this task, we consider that preservation of the tone is mandatory whenever possible. Therefore, we decided to change the tag for the word SARCASMO: `\s -> SARCASMO`.

Anonymize Reddit Reference Reddit usernames, usually found in the form `/u/username` or `u/username` were replaced by the word USUARIO; subreddit mentions, in the form `/r/subreddit` or `r/subreddit` were replaced by SUBREDDIT; Twitter-like usernames, in the form `@username` were replaced by USUARIO; lastly, for any other possible reference the default was defined as OTRA REFERENCIA.

Convert Emojis to Spanish We used the `emoji` module for Python to, first, convert the emojis to English text; then, we defined a dictionary to map each English emoji translation to Spanish. Some examples are:

- confused face: cara confundida
- worried face: cara preocupada
- slightly frowning face: cara ligeramente fruncida
- frowning face: cara fruncida
- face with open mouth: cara con boca abierta
- hushed face: cara callada
- astonished face: cara asombrada

Limit Repeated Characters Is common use in social networks to use repeated characters as a way to intensify the corresponding expression. Therefore, for each word that presents repeated characters we trimmed the repetitions to two. Example: `puuuuutooo -> puutoo`.

Normalize Whitespaces Whitespaces are mostly used as a formatting element but they do not add nor modify the semantics of the message, therefore a normalization of whitespaces has been carried on with the following steps:

- Replace various types of whitespace with standard space
- Replace multiple spaces with single space
- Remove spaces at the beginning and end of lines
- Limit consecutive blank lines, conserving at most one blank line to delimit paragraphs
- Remove leading and trailing whitespaces for the entire text

3.2. Models Used

Our approach consisted in using different models to generate semantic contextual embeddings for a numeric representation of each text, once the embeddings were generated they were stored in an OpenSearch [26] Vector Knowledge Database; classification was then performed using the k-Nearest Neighbors (kNN) algorithm by embedding the test texts, placing them in the same vector space of the database and assigning the class with the most votes. We decided to use two models for the embeddings:

- Titan Text Embeddings V2 [27], which is a propietary model from Amazon
- BETO [28]

3.2.1. Titan Text Embeddings V2

The Titan model was presented by Amazon on April 2024, it is a model specifically designed to generate text embeddings, it is a multilingual model with support for over 100 languages, including Spanish. Since this model is propietary no fine tuning can be performed, therefore, after preprocessing, texts were directly converted into vector embeddings. Performance validation was carried on with the complete dev set.

3.2.2. BETO

For BETO, we fine tuned the model with the training dataset; the dev set was partitioned equally into two subsets. The first subset monitored the fine tuning process, while the second subset evaluated the performance of the kNN classifier. After fine tuning, we froze the best performance weights and extracted contextual embeddings from the `[CLS]`, since it provides sentence-level representations of the texts, for both the training set and the first partition of the dev set. These embeddings were indexed in the OpenSearch vector database. Finally, we used the second partition of the dev set to perform classification via kNN search over the stored embeddings.

3.3. Classification using kNN

The HOMO-LAT competition asked for the models to classify a text into one of three polarity labels: positive (POS), neutral (NEU) and negative (NEG). In addition to the texts, the dataset provided the country of the subreddits where the text was written and a keyword, a reference to which segment of the LGBT+ community the text was talking about.

Considering that slang and expressions change between countries, even when in all these countries the prevalent language is Spanish, we decided to further filter the search over the vector database by country and keyword on top of the semantic similarity search usually performed, allowing for a more fine grained classification.

In Listing 1 we present the information stored in the OpenSearch database. The size of the embedding depends on the model: 1536 for Titan and 768 for BETO, the metric used for semantic search is Cosine Similarity; additional to the embedding, we also stored the country and keyword of the text (for filtering the search), the sentiment which corresponds to the ground truth's polarity from the dataset and the preprocessed text.

kNN classification was enhanced with filters, for Task 1 we restricted vectors used for classification to those that had either the same keyword (segment of the LGBT+ community) or the same country (from which the text came from).

```
{
  "mappings": {
    "properties": {
      "embedding": {
        "type": "knn_vector",
        "dimension": 1536 (for Titan) or 768 (for BETO),
        "method": {
          "name": "hnsw",
          "space_type": "cosinesimil",
          "engine": "nmslib",
          }
        }
      },
      "country": {
        "type": "keyword"
      },
      "keyword": {
        "type": "keyword"
      },
      "metadata": {
```

```
        "type": "object"
      },
      sentiment": {
        "type": "keyword"
      },
      "text": {
        "type": "text"
      }
    }
  }
}
```

Listing 1: OpenSearch Index Configuration

4. Results

Table 1 shows the F1-score macro average for both our proposals in Task 1, it additionally shows the Precision and Recall values for the BETO model since it was the best ranked from both and, hence, reported by the HOMO-LAT staff. It is clear that the fine tuning of BETO was useful in getting embeddings with better contextual representations for the task, since it overpasses Titan by almost 0.1, a model that did not have any fine tuning but was just trained with general purpose and multilingual support in mind.

Table 1
Results for the HOMO-LAT Task 1 competition for both models: Titan and BETO model on the testing set.

Model	F1-score	Precision	Recall
BETO	0.4661	0.4762	0.4641
Titan	0.3792	N/A	N/A

The moderate performance scores obtained by both models can be attributed to several factors inherent to the dataset composition and the challenging nature of the classification task. An analysis of the dataset reveals that the neutral (NEU) class was significantly overrepresented in the training and development datasets, with 3174 and 888 samples respectively, creating a class imbalance that influenced the model's decision-making process. For the positive (POS) class the number of samples were 633 for the training dataset and 80 for the development set, while the positive (NEG) class had 1960 samples in the training dataset and 475 in the development dataset. It is important to note that in our proposal we decided to run data augmentation by combining both the training and development dataset into a single one before performing classification in the test dataset. Therefore, the class distribution on our custom training set is:

- NEU: 4062 samples
- POS: 713 samples
- NEG: 2435 samples

This imbalance led to a systematic bias where both models, particularly BETO, exhibited a tendency to classify ambiguous or borderline cases as neutral, resulting in numerous false positive predictions for the NEU class. The higher precision (0.4762) compared to recall (0.4641) for the BETO model suggests that while the model was reasonably confident in its positive predictions, it struggled to capture all instances of the minority classes due to the overwhelming presence of neutral samples.

5. Conclusions

This work presents an approach for polarity detection in LGBT+ related content from Latin American Reddit posts, combining retrieval augmented classification with transformer architectures and vector

knowledge databases. Our methodology demonstrates the feasibility of using preprocessed text embeddings stored in OpenSearch vector databases for k-Nearest Neighbors classification, particularly when enhanced with country and keyword filtering to account for regional linguistic variations.

The preprocessing pipeline proved crucial for handling the inherent challenges of social media text, including variable post lengths, informal writing styles, slang usage, and formatting artifacts. Our systematic approach to removing URLs, quote markers, markdown elements, and normalizing emojis to Spanish text representations significantly improved the quality of the input data for downstream processing.

The comparative analysis between Amazon's Titan Text Embeddings V2 and the fine-tuned BETO model reveals the importance of domain-specific adaptation. The fine-tuned BETO model achieved an F1-score of 0.4661, substantially outperforming the general-purpose Titan model (F1-score: 0.3792) by approximately 0.1 points. This performance gap underscores the value of task-specific fine-tuning for contextual embeddings, even when compared to multilingual models with broader language support. The integration of additional metadata filtering by country and keyword within the vector similarity search represents a novel contribution to the field, acknowledging that expressions and slang vary significantly across Latin American countries despite sharing Spanish as the prevalent language. This approach enables more fine-grained classification that considers both semantic similarity and regional linguistic characteristics.

Despite these methodological contributions, the overall F1-scores obtained by our models and similar results reported by other competitors indicate significant challenges inherent to this task. The moderate performance levels suggest that working with limited training data and cross-dialect classification presents substantial difficulties for current NLP approaches. The scarcity of labeled data, particularly for underrepresented dialects and nuanced expressions of polarity toward LGBT+ communities, constrains the ability of transformer models to learn robust representations. Furthermore, the cross-dialect nature of Track 2, where models trained on data from Argentina, Chile, Colombia, and Mexico were evaluated on texts from fifteen different Latin American countries, highlights the complexity of generalizing across regional linguistic variations. To address these limitations, future work should focus on data augmentation techniques specifically designed for dialectal variations, exploration of few-shot learning approaches that can better leverage limited training examples, and the development of transfer learning strategies that can more effectively bridge the gap between source and target dialects. Additionally, incorporating external linguistic resources and dialect-aware pre-training could potentially improve model robustness when faced with the inherent data sparsity and cross-dialectal challenges characteristic of this domain.

Declaration on Generative AI

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

Acknowledgments

The authors gratefully acknowledge the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP under Grant 20230140, Centro de Investigación en Computación) and the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) for their economic support to develop this work.

References

[1] R. Rini, E. Utami, A. D. Hartanto, Systematic literature review of hate speech detection with text mining, in: 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), 2020, pp. 1–6. doi:10.1109/ICORIS50180.2020.9320755.

[2] S. Agarwal, C. R. Chowdary, Combating hate speech using an adaptive ensemble learning model with a case study on covid-19, Expert Systems with Applications 185 (2021) 115632. URL: https://www.sciencedirect.com/science/article/pii/S0957417421010265. doi:https://doi.org/10.1016/j.eswa.2021.115632.
[3] F. Alkomah, X. Ma, A literature review of textual hate speech detection methods and datasets, Information 13 (2022). URL: https://www.mdpi.com/2078-2489/13/6/273. doi:10.3390/info13060273.
[4] K. I. Fredriksen-Goldsen, H.-J. Kim, S. E. Barkan, A. Muraco, C. P. Hoy-Ellis, Health disparities among lesbian, gay, and bisexual older adults: Results from a population-based study, American journal of public health 103 (2013) 1802–1809.
[5] K. I. Fredriksen-Goldsen, L. Cook-Daniels, H.-J. Kim, E. A. Erosheva, C. A. Emlet, C. P. Hoy-Ellis, J. Goldsen, A. Muraco, Physical and mental health of transgender older adults: An at-risk and underserved population, The Gerontologist 54 (2014) 488–500.
[6] M. Shahiki-Tash, J. Armenta-Segura, Z. Ahani, O. Kolesnikova, G. Sidorov, A. Gelbukh, Lidoma at homomex2023@ iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), 2023.
[7] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed toowards the mexican spanish speaking lgbtq+ population, Natural Language Processing 71 (2023).
[8] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, T. Alcántara, M. Soto, C. Macias, Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population, Natural Language Processing 73 (2024).
[9] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.
[10] G. Bel-Enguix, H. Gómez-Adorno, S. Ojeda-Trueba, G. Sierra, J. Barco, E. Lee, J. Dunstan, R. Manrique, Overview of HOMO-LAT at IberLEF 2025: Human-centric polarity detection in Online Messages Oriented to the Latin American-speaking lgbtq+ populaTion, Procesamiento del lenguaje natural 75 (2025) –.
[11] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.
[12] F. Balouchzahi, H. L. Shashirekha, G. Sidorov, Hssd: Hate speech spreader detection using n-grams and voting classifier., in: CLEF (Working Notes), 2021, pp. 1829–1836.
[13] E. Roberts, Automated hate speech detection in a low-resource environment, Journal of the Digital Humanities Association of Southern Africa 5 (2024).
[14] A. Ribeiro, N. Silva, Inf-hateval at semeval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 420–425.
[15] M. Siino, E. Di Nuovo, I. Tinnirello, M. La Cascia, et al., Detection of hate speech spreaders using convolutional neural networks., in: CLEF (Working Notes), 2021, pp. 2126–2136.
[16] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, S. Villata, A multilingual evaluation for online hate speech detection, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–22.
[17] P. Liu, W. Li, L. Zou, Nuli at semeval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 87–91.
[18] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, Hatebert: Retraining bert for abusive language detection in english, arXiv preprint arXiv:2010.12472 (2020).
[19] C. Macias, M. Soto, T. Alcántara, H. Calvo, Impact of text preprocessing and feature selection on hate speech detection in online messages towards the lgbtq+ community in mexico, in: Proceedings

of the Iberian Languages Evaluation Forum (IberLEF 2023), 2023.
[20] E. Rivadeneira-Pérez, M. de Jesús García-Santiago, C. Callejas-Hernández, Cimat-nlp at homomex2023@ iberlef: Machine learning techniques for fine-grained speech detection task (2023).
[21] M. G. Yigezu, O. Kolesnikova, G. Sidorov, A. Gelbukh, Transformer-based hate speech detection for multi-class and multi-label classification (2023).
[22] C. F. Rosauro, M. Cuadros, Hate speech detection against the mexican spanish lgbtq+ community using bert-based transformers (2023).
[23] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).
[24] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, arXiv preprint arXiv:2111.09453 (2021).
[25] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).
[26] OpenSearch Contributors, Opensearch: Open source search and analytics suite, 2021. URL: https://opensearch.org/.
[27] Amazon Web Services, Titan embed text v2, 2023. URL: https://us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/model-catalog/serverless/amazon.titan-embed-text-v2:0.
[28] 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.