HomoCIC at HOMO-MEX 2024: Deep Learning Approaches for Classifying Homophobic Content in Tweets and Songs: Leveraging LLM and NL

Omar Garcia Vazquez[1,†], Marco Cardoso-Moreno[1,*,†], José Alberto Torres-León[1,†] and Diana Jiménez[1,†]

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

Abstract

The increase of use of social media platforms has led to an increase in hate speech expressions in these platforms, including homophobic content targeting the LGBT+ community. Since LGBT+ people presents a particular susceptibility to various forms of discrimination and mental health issues, the wide spread of hate speech expressions poses a significant societal risk, particularly in the context of Mexican society, where drug abuse presents a pervasive social challenge.

The HOMO-MEX task, par of the IberLEF (Iberian Languages Evaluation Forum), aims to tackle this issue by developing Natural Language Processing systems to detect hate speech directed to the LGBT+ community. The 2024 edition introduced three tracks: a multi-class hate speech detection, a multilabel one and, for the first time, a track focusing on identifying homophobic hate speech content in song lyrics.

Our proposal consists on the use of different Large Language Models, namely: BERT, DistillBERT and RoBERTa, for tracks one and three, achieving 0.8219 and 0.4896 of macro F1-score, respectively. Our findings demonstrate the effectiveness of these advanced computational techniques in identifying subtle expressions of hate speech, contributing to the broader effort of mitigating the spread of harmful content and fostering a safer online and cultural environment for LGBT+ communities.

Keywords

Hate Speech, Homophobia, LLM, NLP, Classification

1. Introduction

In recent years there has been an increase in hate speech expressions, mainly due to the increase in social media activity [1]. According to the European Union, hate speech is defined 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].

Since hate expressions consist of offensive and harmful content targeting specific communities, they are prone to cause harm and conflict; furthermore, it is easy for such expressions to be widely spread due to prejudices [3]. In particular, homophobic hate speech is of particular importance, given that LGBT+ members suffer from substance abuse disorders, mental health issues, job market discrimination, as well as limited access to health care services [4, 5]. This fact is even more important under the context of the Mexican society, where drugs consumption presents a social issue that do not only affects the LGBT+ community [6].

It is, under this context, that the HOMO-MEX task [7, 8] arises as part of the IberLEF (Iberian Languages Evaluation Forum) [9]. The main objective of the task is the development of Natural

IberLEF 2024, September 2024, Valladolid, Spain

*Corresponding author.

†These authors contributed equally.

ogarciav2024@cic.ipn.mx (O. G. Vazquez); mcardosom2021@cic.ipn.mx (M. Cardoso-Moreno); jtorresl2019@cic.ipn.mx (J. A. Torres-León); dianaljl.99@gmail.com (D. Jiménez)

https://www.linkedin.com/in/omar-garcia-vazquez-093128219/ (O. G. Vazquez); https://cardoso1994.github.io/ (M. Cardoso-Moreno); https://github.com/JAlbertoTorres (J. A. Torres-León)

0009-0001-4391-6225 (O. G. Vazquez); 0009-0001-1072-2985 (M. Cardoso-Moreno); 0000-0003-2704-0216 (J. A. Torres-León); 000000023326557X (D. Jiménez)

Language Processing (NLP) systems that are able to identify, in Spanish written tweets, LGBT+ related hate speech no matter how subtle the expression. In the 2024 edition of HOMO-MEX [8] there were three tracks: a multi-class Hate speech detection track, where tweets are mapped to three classes, LGBT+phobic, not LGBT+phobic and not LGBT+related; a multilabel hate speech detection track where the possible classes are Lesbophobia, Gayphobia, Biphobia, Transphobia, Other LGBT+phobia and Not LGBT+related. Lastly, track 3 consisted on the classification of song lyrics containing LGBT+phobic hate speech, where classes were defined as LGBT+phobic and Not LGBT+phobic; it is the first time that the HOMO-MEX task includes this track, arguing on the difficulty to identify hate speech in songs, since this detection depends on the context and culture under which the songs were written.

The rest of this manuscript is structured as follows: Section 2 presents a brief literature review for hate speech detection, at first instance as a general overview and in second place specific to the HOMO-MEX task; Section 3 explains our proposal, including preprocessing, models and metrics; Section 4 shows the results obtained; finally, Section 5 highlights the importance of our proposal and points out to future directions for subsequent research.

2. Literature Review

In this section, a brief literature review on hate speech and homophobic content is presented. First in general terms for the task and, finally, specific to the HOMO-MEX task.

2.1. General Overview

Traditional Machine Learning (ML) algorithms, in conjunction with NLP preprocessing techniques. For instance, in [10] used a voting classifier conformed of: Support Vector Machine (SVM), Random Forest (RF) and Logistic Regression (LR), together with character and word n-grams, and syntactic ngrams; the model was developed for the Profiling Hate Speech Spreaders (HSSs) task from PAN at CLEF 2021, achieving accuracies of 0.73 and 0.83 for English and Spanish. Additionally, in [11] three tree-based algorithms were used, namely RF, Light Gradient Boosting Machine (LightGBM) and Cat Boost classifiers, all with bayesian optimization and unigrams and bigrams as features, achieving accuracy scores from 0.85 to 0.87 depending on the model used.

Convolutional Neural Networks (ConvNets) have been extensively used for hate speech detection. Ribeiro and da Silva [12] proposed a ConvNet for hate speech classification for the SemEval-2019 Task 5; the model used pre-trained word embeddings such as GloVe and FastText with 300 dimensions, achieveing F1-scores between 0.48 and 0.69. Siino and colleagues [13] worked on the HSSs task, they used a ConvNet with a single convolutional layer getting an accuracy of 0.79 on a multilingual (English and Spanish) setup, while achieving individually accuracy values of 0.85 and 0.73 for Spanish and English, respectively.

In [2], they used the A-stacking classifier, based on ensemble learning; it uses a Recurrent Neural Network (RNN) to create word embeddings, a Long Short-Term Memory (LSTM) network and softmax activation to detect hate speech across several dedicated datasets, both in within datasets and cross datasets setups. Corazza and colleagues [14] proposed a modular neural network consisting on an RNN layer, a dense layer of 100 neurons and a single output neuron; this model is able to support both word-level and tweet-level features, it was tested on English, German and Italian languages.

The NULI team at SemEval-2019 [15] fine tuned the Bidirectional Encoder Representations from Transformers (BERT) to detect hate speech which made the team get the first place at the competition; preprocessing consisted only on emoji substitution, hashtag segmentation and converting all text to lowercase.

Lastly, Caselli et al. [16] introduced HateBERT, a re-trained BERT specifically for abusive language in English; trained on the Reddit Abusive Language English datase (RAL-E). After training, HateBERT was tested—along with the original BERT model—in several datasets, as a consequence of this re-training, HateBERT outperformed BERT in all benchmarks.

2.2. HOMO-MEX Literature Review

The year 2023 marks the first iteration of the HOMO-MEX [7] task at IberLEF. A Mexican Spanish tweets corpus containing nouns indicative of the LGBT+ community was created for Hate Speech detection. The selection of slang, slur, general terminology and nouns was performed by a list of words from social network channels like Twitter, Facebook and Instagram, among others. Variations in terms were also considered, for instance, the meaning of the word *puto* can also be expressed by the words: *pute* and *putx* (efeminitaztion); *putito, putín* (diminutive); and *putote, putón* (augmentative); to name a few examples.

After the terms extraction, web scrapping over time was performed to extract 706,886 tweets in Mexican Spanish, from those, 11,000 were annotated into LGBT+phobic, not LGBT+phobic and not relevant to the LGBT+ community; additionally, in the multilabel approach tweets could belong to one or more of the following categories: Gayphobia, Lesbophobia, Biphobia, Transphobia or other types of LGBT+phobia.

Among the proposals presented in this first edition, [17] used traditional NLP preprocessing and feature extraction techniques, such as Bag of Words (BoW) and term frequency (TF), and the inverse document frequency (IDF) to perform TF-IDF. As classifiers they used a Linear Support Vector Machine (LSVM) and a Bagging Classifier for with LSVM as base model. Rivadeneira and colleagues [18] participated only in the second (multilabel) task. They trained individual classifiers for each LGBT+phobia class; as features they used n-grams for word tokens and a weighted TF-IDF BoW representation. For classifiers they used Random Forest and SVMs.

In [19] several transformer-based models such as BERT and RoBERTa proved to be effective for tracks 1 and 2, detecting LGBT+phobic content in multiclass and multilabel setups, respectively. Similarly, in [6] BERT-based models were also used for the first track, achieving a Macro F1 score of 0.73. Additionally, Rosauro and Cuadros [20] used BETO [21], RoBERTuito [22] and mDeBERTa [23] (all BERT based models) for tracks 1 and 2, achieving 0.84 and 0.68 of Macro F1 score, respectively.

3. Proposal

This section provides insight on our proposal for tracks 1 and 3 of the 2024 edition of HOMO-MEX [8].

3.1. Preprocessing

Based on the conclussions we draw from our state of the art review, we decided to work with transformer based models only. Therefore, our preprocessing is minimal, consisting of the following steps: remove URLs, remove Hashtags, remove Twitter handles—marked by the @ symbol—, and, lastly, remove emojis.

3.2. Models Used

The models we decided to work with were all transformer based and downloaded from the HugginFace repository. The selected models were:

- DistillBERT [24] multilingual cased,
- RoBERTa [25] base, and
- BERT [26] multilingual cased

4. Results

For experimentation, we used Hold-out validation splits over the training set: 80% was used for training and 20% was used for validation of the transformer based models. Once we considered the training phase as finished, we passed the testing set to the models. In Table 1 we show the results of the several models for track 1: multi-class hate speech detection.

Table 1
Macro F1 score results over the testing set on Track 1 for the selected models.

Model	Macro F1 score
RoBERTa	0.8219
BERT	0.8219
DistillBERT	0.7892

Additionally, in Table 2, we show the results for track 3—classification of song lyrics containing LGBT+phobic hate speech—for the RoBERTa model, which showed the best results on track 1.

Table 2
Macro F1 score results over the testing set on Track 3 for the BERT multilingual cased model.

Model	Macro F1 score
RoBERTa	0.4896

5. Conclusions

In this study, we have explored the efficacy of various Large Language Models (LLMs) in classifying homophobic content in tweets and songs, contributing to the HOMO-MEX task under the IberLEF initiative. Our research demonstrates the potential and limitations of different transformer-based models, specifically BERT, DistillBERT, and RoBERTa, in detecting hate speech targeted at the LGBT+ community.

Our findings show the robustness of transformer based models, in conjunction with fine tuning procedures, when used for context-dependet hate speech. RoBERTa and BERT both achieved a Macro F1 score of 0.8219 in the multi-class hate speech detection track.

The performance of RoBERTa in identifying homophobic content within song lyrics was notably lower, achieving a Macro F1 score of 0.4896, thus, highlighting the complexity of the challenges present in lyrical content, which often contains metaphorical or figurative language, cultural references and style variations within writers, making it difficult for the models to detect hate speech.

The use of advanced NLP models—by flagging harmful content—contributes significantly to the efforts being made by governments and digital platforms to create safer environments for the LGBT+ community.

Our study confirms that leveraging LLMs in hate speech detection is a promising approach, yet it also highlights the need for continuous refinement and adaptation to address the evolving landscape of online and cultural expressions of hate.

Acknowledgments

The authors gratefully acknowledge the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP under Grant 20230140, Centro de Investigación en Computación) and the Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) 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. G'omez-Adorno, G. Sierra, J. V'asquez, 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. G'omez-Adorno, G. Bel-Enguix, H. Calvo, J. V'asquez, S. T. Andersen, S. Ojeda-Trueba, T. Alc'antara, 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] 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.
[11] E. Roberts, Automated hate speech detection in a low-resource environment, Journal of the Digital Humanities Association of Southern Africa 5 (2024).
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, Hatebert: Retraining bert for abusive language detection in english, arXiv preprint arXiv:2010.12472 (2020).
[17] 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.
[18] 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).
[19] M. G. Yigezu, O. Kolesnikova, G. Sidorov, A. Gelbukh, Transformer-based hate speech detection for multi-class and multi-label classification (2023).
[20] C. F. Rosauro, M. Cuadros, Hate speech detection against the mexican spanish lgbtq+ community using bert-based transformers (2023).
[21] 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).
[22] 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).
[23] 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).
[24] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. arXiv:1910.01108.
[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.
[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.