The LaboCIC at HOMO-MEX 2024: Using BERT to Classify Hate-LGTB Speech

Daniel Yacob Espinosa, Grigori Sidorov and Eusebio Ricárdez Vázquez

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico

Abstract

Social media has emerged as a crucial space for communication, especially with the increase in its use during the pandemic. These platforms enable the exchange of information and connection among users, being particularly significant for communities such as the LGBT+ movement. However, cyberbullying towards the LGBT+ community on social media has severe consequences, including psychological harm, isolation, low self-esteem, and in extreme cases, physical violence and even deaths. Despite the policies and tools implemented by platforms to combat hate, their effectiveness varies, and the LGBT+ community remains highly exposed to these behaviors.

One of the current challenges in content moderation on social media is identifying satire and irony, complicating the classification of messages as hate content. In this context, we participated in the tasks proposed by Homo-Mex [1], focusing on Task 1 and Task 3. Task 1 centers on the classification of tweets with hate content directed at the LGBT+ community, while Task 3 involves binary classification of songs in Spanish. To solve these problems, we used BERT, achieving results of **92.37%** F1-score for Task 1 and **89.14%** F1-score for Task 3. This research work aims to improve artificial intelligence systems for the categorization of hate speech, contributing to the creation of safer digital spaces for the LGBT+ community.

Keywords

BERT, tweets, LGBT+, hate speech, hate LGBT+

1. Introduction

Social media has become a fundamental space for communication, and since the pandemic in recent times, the use of these digital spaces has been increasing. This is due to the exchange of information among users and their interactions. Many of these users find social media to be a safe place to connect with communities such as the LGBT+ movement. The term LGBT+ is a way to recognize and group people who identify with sexual orientations different from heterosexuality and cisgender. However, on many digital platforms, social biases still exist, which divide the population with this type of content. Cyberbullying towards the LGBT+ community on social media has serious consequences, both for the individuals who are directly attacked and for the community as a whole. Such acts can lead to psychological harm, feelings of isolation, low self-esteem, and in extreme cases, situations of physical violence and intimidation, with even some cases resulting in fatalities [2]. Social media platforms have implemented various policies and tools to combat the spread of hate, such as content moderation, reporting systems, and the promotion of safe spaces, although their effectiveness and consistency can vary. These measures are implemented for all social media users, but due to the impact and large community of the LGBT+ movement, these behaviors are more exposed [3].

Research by Andrew Flores indicates that the LGBT+ community is more likely to be victims of hate crimes and discrimination. These crimes are primarily motivated by social prejudices, and when members of these communities seek help, very few countries and institutions offer specialized support for this type of aggression [4]. It is noted that the highest incidence of these crimes occurs in close circles: schools, workplaces, or the homes of the affected individuals, and that the harm caused by these crimes is more severe and violent than other incidents.

IberLEF 2024, September 2024, Valladolid, Spain

espinosagonzalezdaniel@gmail.com (D. Y. Espinosa); sidorov@cic.ipn.mx (G. Sidorov); ericardez@cic.ipn.mx (E. R. Vázquez)

http://www.cic.ipn.mx/~sidorov/ (G. Sidorov)

0009-0004-9245-2350 (D. Y. Espinosa); 0000-0003-3901-3522 (G. Sidorov)

There is an ongoing problem that is still under investigation: satire and irony. Many of the messages displayed on social media have this aspect, making it more difficult to classify them as hate content towards certain individuals or communities [5].

For this occasion, we decided to participate in the task created by Homo-Mex [1]. Homo-Mex, for this year, decided to launch three research tasks for this year is IberLEF [6]. For this research work, we chose to work on Task 1 and Task 3. All the tasks are related to hate speech towards the LGBT+ community; given the issues that such behaviors can cause, Homo-Mex is tasks aim to improve artificial intelligence systems for categorizing these topics.

In Task 1, the objective is to classify tweets with hate content directed at the LGBT+ community. For Task 2, within a set of tweets, a marker is placed to identify the specific type of community the hate is directed towards: lesbophobia, gayphobia, biphobia, transphobia, LGBT+phobia, or not LGBT+ related. For the final task, Task 3, the goal is to perform binary classification for a set of songs in Spanish. In this research work, we will focus only on Task 1 and Task 3.

2. Dataset

In Task 1, the objective is to classify tweets with hate content directed at the LGBT+ community, using a dataset composed of 7,000 tweets in Spanish for training and 220 tweets for testing. These tweets exhibit typical characteristics of social media texts, such as mentions of other users, hashtags, links, and emojis. To address this task, various preprocessing techniques were implemented to clean and organize the text. Subsequently, natural language processing (NLP) models, particularly BERT and its variations, were applied to identify patterns and perform the classification.

In Task 3, the training dataset consisted of 600 Spanish songs, and the testing dataset comprised 246 songs. Most of these songs are segmented according to their musical structure into parts such as intro, chorus, verses, bridge, refrains, and outro. We considered it important to implement a preprocessing layer, as these elements are not deemed significant enough to be used in the research.

Table 1
Dataset of HomoMex 2024[1]

Dataset	Train	Test
Task 1	7000 tweets	220 tweets
Task 3	600 songs	246 songs

3. Metodology

We conducted several experiments with tweets and found that in most cases, we recommend adding a preprocessing layer to the data. Since Tasks 1 and 3 have different characteristics, we implemented different preprocessing approaches for each dataset. For Task 1, we have tweets, which, as we know, often include informal language, abbreviations, and emojis. In contrast, Task 3 involves songs, typically filled with informal language and slang, which can be mixed with metaphors, ambiguities, or even vulgar expressions disguised with other texts.

We will start with preprocessing layers before feeding the data into the models, ensuring that our entire research is related to BERT and some of its variations.

3.1. Pre-processing steps

As with any text-related task, we always recommend using a preprocessing layer before directly applying models. Preprocessing the data helps transform, clean, and prepare it for analysis or modeling. This improves data quality and makes it easier to use in algorithms. Additionally, preprocessing helps enhance model performance and facilitates the interpretation and analysis of the results. Without a preprocessing layer, the outcomes of an analysis or machine learning model can be inaccurate or misleading.

The following configuration was used exclusively for the tweets in Task 1:

Lowercase All tweets were converted to lowercase to standardize the texts.

Links Links were replaced with the tag 'enlace'.

Hashtags Hashtags were modified and replaced with the tag 'hashtag'.

User Mentions User mentions were replaced with the tag 'mención de usuario'.

Emojis No changes were made to emojis as they integrate well with the model.

Other Symbols All symbols not recognized within the ASCII reference standard were removed.

Due to Task 3 involving songs throughout its dataset, the following configurations were made:

Lowercase All texts were converted to lowercase.

Musical Structures Marks of musical structures were removed.

Other Symbols All symbols not recognized within the ASCII reference standard were removed from the songs.

4. Experiments

In previous works, we have been involved in tweet classification, primarily focusing on bots [7]. Therefore, we wanted to test our methodology to observe its behavior with a different task. In this case, we used a structure of word and character N-grams. The tests conducted with this structure were for both tasks.

Table 2
Results **Task 1** of F1-score with **N-grams Structure**

N-grams char	word	F1-Score
5-7-8-9	3-4-5	75.21

Table 3
Results **Task 3** of F1-score with **N-grams Structure**

N-grams char	word	F1-Score
5-7-8-9	3-4-5	66.30

In Task 1, focused on classifying tweets with hate content, we achieved expected results, demonstrating the effectiveness of N-grams in this context, although there was still significant room for improvement. In Task 3, which involved the binary classification of songs in Spanish, N-grams also proved to be useful; however, the challenge was greater due to the different nature of the texts and the more creative and varied use of language in song lyrics. Thanks to the previous results, which were far from accurate, we decided to try another methodology that we have been using in recent years.

For PAN 2023, we conducted research on Crypto-influencers classification using BERT [8]. In our experiments for PAN 2023, the standout models were BERT, RoBERTa [9], and BERTweet [10]. Therefore,

we decided to test which of these models we could use for this research. It is important to mention that for all models used in the experiments, we applied the following configuration: a batch size of 16, 9 training epochs, and a GPU usage limit of 15GB. We used PyTorch with pre-trained models on the GPU. These configurations were applied to all three BERT variants and evaluated using F1-score.

Table 4
Results of **Task 1 F1-Score** and **BERT Variations**

Model	F1-Score
RoBerta	87.27
BERTweet	88.41
BERT	92.37

For the next Task, in our case Task 3, we decided to conduct the same experiment with all three BERT variants. These were the results obtained for this task.

Table 5
Results of **Task 3 F1-Score** and **BERT Variations**

Model	F1-Score
RoBerta	83.99
BERTweet	74.02
BERT	89.14

For these experiments, we consider utilizing more GPU power if necessary when using RoBERTa, as this model resulted in significantly longer training times. Surely, with additional computational resources, we could potentially leverage its robustness more effectively.

5. Conclusions

We believe that these types of problems are currently underexplored in research, despite their great importance. Additionally, something we greatly appreciated about this research was the datasets, as they were all in Spanish, whereas most models are trained and built for the English language. Often, the problem lies in the lack of Spanish data to train these models.

For Task 1, we were surprised by the results of BERTweet. Although this model typically performs well, given its primary use with tweets, an important aspect is that it was trained on English tweets. This led us to reflect on the importance of having robust models available and trained in multiple languages. RoBERTa also showed promising results, but its performance in Spanish was noticeably inferior, further emphasizing the need for models specifically trained for different languages.

For Task 3, the most notable aspect is that we used the same models as for Task 1. Despite being different data sets, they showed similar results. We suppose that this is why BERtweet performance does not show more outstanding results, mainly due to its training method and the fact that the data were in English, lacking the necessary approach to achieve good classification results. This data offers us valuable lessons on the importance of training data in the effectiveness of artificial intelligence models. The variability in the data and the adaptation of the model to different languages and types of content are crucial factors that affect performance.

These experiments provided a unique opportunity to evaluate the flexibility and adaptability of our models in different linguistic and thematic contexts. It also allows us to better understand the limitations of current models and highlights the need to develop more robust training techniques that consider the particularities of language, the context of the data, and the cultural context.

These findings led us to pose several additional research questions. How can we improve language models to perform equally well in Spanish as they do in English? What strategies can be adopted to generate and curate more Spanish datasets to enable more effective training of these models?

We would like to contribute to the creation of a model similar to BERTweet, trained exclusively with large-scale Spanish tweets. This would not only improve the classification of bots and the detection of

fake news on Spanish-speaking social networks but could also be applied to other natural language processing tasks, such as hate speech in different internet communities.

References

[1] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, T. Alcáhntara, 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).

[2] Z. Akmeşe, K. Deniz, Hate speech in social media: Lgbti persons, 2017.

[3] M. A. Walters, J. Paterson, R. Brown, L. McDonnell, Hate crimes against trans people: Assessing emotions, behaviors, and attitudes toward criminal justice agencies, J Interpers Violence 35 (2017) 4583–4613.

[4] A. Flores, R. Stotzer, I. Meyer, L. Langton, Hate crimes against lgbt people: National crime victimization survey, 2017-2019, PLOS ONE 17 (2022) e0279363. doi:10.1371/journal.pone.0279363.

[5] W. Yu, B. T. Boenninghoff, D. Kolossa, Bert-based ironic authors profiling, in: Conference and Labs of the Evaluation Forum, 2022. URL: https://api.semanticscholar.org/CorpusID:251471104.

[6] 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.

[7] D. Espinosa, H. Gómez-Adorno, G. Sidorov, Bots and Gender Profiling using Character Bigrams, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[8] D. Y. Espinosa, G. Sidorov, Using BERT to profiling cryptocurrency influencers, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2568–2573. URL: https://ceur-ws.org/Vol-3497/paper-207.pdf.

[9] 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.

[10] D. Q. Nguyen, T. Vu, A. Tuan Nguyen, BERTweet: A pre-trained language model for English tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 9–14. URL: https://aclanthology.org/2020.emnlp-demos.2. doi:10.18653/v1/2020.emnlp-demos.2.