UC-CUJAE at HOMO-MEX 2024: Detecting Hate Speech Against the LGTB+ Community using Transformers on Imbalanced Datasets

Anibal Hernández-González [1], Julio Madera-Quintana [1*] and Alfredo Simón-Cuevas [2]

[1] *University of Camagüey, Circunvalación Norte km 5 1/2, Camagüey, Cuba*
[2] *Universidad Tecnológica de La Habana José Antonio Echeverría, Marianao, La Habana, Cuba*

Abstract

The pervasive hate manifestations in social communication spaces and culture pose a significant challenge for contemporary societies. The sheer volume of daily information exacerbates the difficulty of detecting aggressive content targeting specific groups. The LGBTQ+ community is disproportionately affected by this problem. In order to promote a more positive and inclusive environment for the LGBTQ+ community, the Homo-Mex 2024 shared task proposes the development of automated learning systems capable of tackling various subtasks to create safer and healthier online spaces for the LGBTQ+ community. This paper proposes using transformer-based techniques to present solutions to the three problems posed at the event: multi-class, multi-label classification, and binary classification. The results show that the proposed procedure effectively addresses the undertaken tasks, with the binary classification yielding particularly noteworthy outcomes.

Keywords

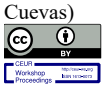
Imbalanced Classification, Transformers, Data Augmentation, LGBT-Phobia, Hate Speech Detection

1. Introduction

Detecting and classifying hate speech against the LGBTQ+ community in social media and artistic products is crucial for creating safe spaces for community exchange and growth. In Mexico, social media and artistic products often feature comments that target and aggress against this population sector. However, most NLP studies focus on English, leaving the Spanish-speaking community vulnerable [1].

To address this gap, Homo-Mex 2024, at the context of IberLEF 2024 [2] proposes three tracks to develop solutions. Our goal is to develop classification models that can accurately identify hate speech directed towards the LGBTQ+ community and pinpoint the specific subgroup within the community that is being targeted, using multi-label, multi-class, and binary classification approaches [3].

Classification tasks are a common problem in the field of artificial intelligence. For this reason, numerous approaches have been taken to address each problem. While this is an advantage, the challenge presents difficulties in deciding which techniques are the most convenient for tackling each particular problem and how existing methods can be adapted to the particular characteristics of the proposed tracks. The decision to apply or not preprocessing techniques, the approach to dealing with imbalance in the training set, or the selection of classification models to use has a crucial impact on the quality of the results. For this reason, we propose finding solutions that consider all these factors to obtain an alternative that contributes to combating hate speech in the research context [4, 5].

IberLEF 2024, September 2024, Valladolid, Spain
EMAIL: anibal.hernandez@reduc.edu.cu (A. Hernández-González); julio.madera@reduc.edu.cu (J. Madera-Quintana); asimon@ceis.cujae.edu.cu (A. Simón-Cuevas)
ORCID: 0009-0006-5337-7954 (A. Hernández-González); 0000-0001-5551-690X (J. Madera-Quintana); 0000-0002-6776-9434 (A. Simón-Cuevas)


2. Related work

The field of Natural Language Processing (NLP) has made tremendous progress in recent years [6], mainly driven by the emergence of transformer models. Models like BERT, GPT, and RoBERTa have transformed how we process and understand text, allowing us to tackle complex linguistic tasks with unprecedented precision [7, 8]. A critical application of NLP is the identification of hate speech and discriminatory content, mainly when directed towards vulnerable groups such as the LGBTQ+ community [3].

Recent studies have utilized transformer models to identify hate speech targeting the LGBTQ+ community. For instance, one study [9] demonstrated that fine-tuning pre-trained transformer models on annotated LGBTQ+ hate speech datasets substantially enhanced their performance. By harnessing the contextualized representations generated by these models, researchers could pinpoint the subtle language patterns and nuances characteristic of hate speech.

These recent research showcase the potential of transformer-based models in detecting hate messages against the LGBTQ+ community. Researchers have achieved significant advancements in accurately identifying and categorizing discriminatory content by training these models on large, annotated datasets and fine-tuning them precisely for hate speech detection. The use of transformer models has proven instrumental in capturing the intricate linguistic characteristics of hate speech, allowing for more effective moderation of online platforms, the protection of vulnerable communities, and the promotion of a safer and more inclusive digital environment [10]

3. Datasets and tracks

The Corpus provided by the organizers is described at Codalab (https://www.codabench.org/competitions/2229). This Corpus contains three archives per phase (development, training, and test). In each phase we have three dataset, one per task.

3.1. Track1: Hate speech detection track (multi-class)

The objective of this task is to predict the label of each individual tweet (see Table 1 and Table 2). The three possible labels a tweet can have are LGBT+phobic (P), not LGBT+phobic (NP) , and not LGBT+related (NR). We define each one of these labels next:

- LGBT+phobic (P) tweets contain hate speech directed against any person whose sexual orientation and/or gender identity differs from cis-heterosexuality.
- Not LGBT+phobic (NP) tweets are those that do not include any hate speech against the LGBT+ population but do mention this community.
- Not LGBT+related (NR) tweets are those that are not related in any way to the LGBT+ community.

In this first track, participants will be able to assign one label to each tweet, e.g. Tweet X can be assigned the label "P", and Tweet Y can be assigned with label "NR".

3.2. Track 2: Fine-grained hate speech detection track (multi-labeled)

The objective of this task is to predict one or more label(s) of each individual tweet that contains LGBT+phobic hate speech (See Table 3 and Table 4).

The labels in this sub-track are related to various phobias related to LGBT+phobia. Each one of these phobias is described next:

- **Lesbophobia** is homophobia explicitly directed at homosexual people who identify as female.
- **Gayphobia** is homophobia explicitly directed at homosexuals who identify as male.
- **Biphobia** refers to hate speech directed against people who are attracted to more than one gender.

- **Transphobia** refers to hate speech directed against non-cis-gendered people.
- **Other LGBT+phobia** is hate speech against other sexual and gender minorities not included in any of the categories described above (e.g "aphobia" which describes the hatred received by people who do not feel sexual attraction).
- **Not LGBT+related (NR).** tweets are those that are not related in any way to the LGBT+ community.

Table 1
Class distribution for Track 1.

Class	Development	Training
P	862	1072
NP	4360	5482
NR	1778	2246

Some instances for Track 1.

Table 2
Examples of instances for Track 1.

Index	Tweet	Label
107	jajaja, es que las lesbianas me dan miedo :(.	P
23	¡Ven a México! Anda. Aquí hay tremenda fiesta Trans. @Do1erroberto corrobora. Acaba de llegar.	NP
92	Solitario sujeto incendia el bazar La Onza que se ubica en Sarabia 6na-Bis de la zona centro de Cd. Lerdo.	NR

In this second track, participants will be able to assign one or more labels to each tweet, e.g. Tweet X can be assigned the labels "L", "G", "B", and Tweet Y can be assigned only the label "O".

Given that each tweet can have one to five labels, the output submission must follow this strict order: "L", "G", "B", "T", "O", "NR".

The tweets assigned less than five labels must include the string "0" in place of the label(s) that were not predicted for such tweets.

3.3. Track 3: Homophobic lyrics detection track (binary)

The objective of this task is to predict if a phrase of a lyrics song contains LGBT+phobic hate speech. This is a binary task (LGBT+phobic (P), not LGBT+phobic (NP)). We define each one of these labels next:

- **LGBT+phobic (P).** Lyrics contain hate speech directed against any person whose sexual orientation and/or gender identity differs from cis-heterosexuality.
- **Not LGBT+phobic (NP).** Lyrics are those that do not include any hate speech against the LGBT+ population but do mention this community.

In this last track, participants will be able to assign one of the two labels to each tweet, e.g. Tweet X can be assigned the label "P", and Tweet Y can be assigned with label "NP" (See Table 5 and 6).

Table 3
Label distribution for Track 2.

Class	Development	Training
L	**72**	**88**
G	714	894
B	10	10
T	79	94
0	64	77

Table 4
Examples of instances for Track 2.

Index	Tweet	Label
19	Pide a @Disney que no adoctrine a nuestros hijos con ideología de género en sus programas infantiles.	0,0,0,0,1,0
39	Estoy seguro que en lo más profundo del infierno hay un lugar reservado para esas personas que utilizan escopetas en línea.	0,1,0,0,0,0
59	En una sociedad donde las preferencias sexuales son una moda y no una libre elección aparentemente innata, aborrezco a los "homosexuales"	1,1,0,0,0,0

Given that each tweet can have one of these tree labels, the output submission must have two columns, first column corresponds to the ID of the tweet and the second column corresponds to your prediction value for each tweet indicating "P" or "NP".

Table 5
Class distribution for Track 3.

Class	Development	Training
P	**40**	**40**
NP	**560**	**945**

For the test dataset, we do not know the class distribution; we only know the instances to classify. Table 7 shows the number of instances to classify per track.

Table 6
Examples of instances for Track 3.

Index	Tweet	Label
1	Todavía yo te espero Aunque yo sé que tú no vas a volver Todavía yo te quiero	NP
3	No se puede corregir a la naturaleza Palo que nace dobla'o, jamás su tronco endereza Y mientras pasan los años, el viejo cediendo un poco	P

The number of unknown instances to classify per track are shown in Table 7.

Table 7
Number of instances per task to classify in the test dataset.

Track	Instances
Track 1	**2200**
Track 2	**268**
Track 3	**246**

4. Methodology

This study addresses the challenges posed by Homo-Mex 2024 by leveraging the datasets provided by the event organizers, which exhibit distinct characteristics. A thorough examination of each track's requirements informs the development of tailored preprocessing strategies and the judicious application of data augmentation techniques to enhance model performance [11, 12]. A range of innovative models is selected, encompassing traditional classification approaches, state-of-the-art transformer architectures, and influential ensemble methods. Through rigorous cross-validation, we evaluated each approach's performance and identified the most effective combinations of preprocessing, data balancing, and classification systems for each task. This iterative and adaptive approach enables optimizing algorithmic performance, yielding outstanding results. We employ the evaluation metrics recommended by the organizers to ensure the highest level of accuracy, thereby providing a comprehensive and reliable assessment of our methodologies.

4.1. Proposed procedure

Given that each dataset presents such varied characteristics, we analyzed the preprocessing performed on each dataset separately. After conducting several tests for the first track (multi-class classification), it is agreed not to use preprocessing techniques, as they worsen the models' ability to identify relevant features during training. As for data balancing, We decided to ignore it since there is no significant imbalance between the classes assigned to the instances. Finally, different alternatives are tested for classifying the instances, obtaining the best results with a tuned variant of the pre-trained xlm-RoBERTa classification model.

For the second track (multi-class classification), the imbalance problem was more pronounced, so we explored different approaches. The first approach consists of treating each possible combination of labels as a class and applying the techniques used in task 1. The second approach (which ultimately yields better results) is to treat each instance as a set of binary problems and then apply a variant of the method proposed for track 3.

We tested different data augmentation techniques in the different approaches to address the imbalance problem in the training set. None of the balancing techniques used have the desired impact, so the alternative technique of assigning thresholds to the labels is employed to influence the model's predisposition and improve the detection of minority labels. This adjustment proves to be effective in improving the detection of minority labels. In this track, transformer architectures and ensembles, described in section 4.3, are employed and tuned specifically for this proposal. Once again, the best results were obtained using the xlm-RoBERTa architecture. An interesting detail is that the label labeled 'NR' (Not-related) had no instances assigned, so no classifier learned to identify it. To address this problem, and considering that no instance classified as 'NR' can receive another classification, it is determined that if an instance is not classified as any of the other labels, it will receive this classification.

In track 3, our team obtained the most relevant results. Again, different preprocessing techniques are experimented with. The results tend to improve slightly when no preprocessing is used. By eliminating certain elements that are traditionally not considered valuable for natural language analysis, such as prepositions, conjunctions, and others, the context of the expressions was altered. Since the transformer models are context-sensitive, the ability to identify relevant features is lost. Unlike the previous two, applying balancing techniques improves the system's ability to discriminate instances correctly in this track.

Synthetic instances were generated using two approaches to address the imbalance. The first is based on synonyms, using the Wordnet knowledge base to substitute similar terms, and the second uses word-embedding techniques to generate semantically similar instances, employing the BERT architecture. Both variants positively impact the results, with the best solutions obtained through the generation of instances based on transformers. Subsequently, different variants of classifiers are employed, including those previously exposed, with parameters adjusted precisely for this task. For the third time, the best results were obtained with the xlm-RoBERTa architecture, in this case, with a pre-trained model on Twitter hate speech detection that was later trained on the specific dataset for this track. It is worth noting that, in this case, this solution obtains the best results of all proposals in the event.

4.2. Data augmentation

Data augmentation is a technique used to increase the size and diversity of a dataset by applying transformations to existing data generating new synthetic data that can be used to train machine learning models. This helps to improve the model's performance and reduce overfitting.

Regarding data augmentation applied to text, some standard techniques include substitution, insertion, or deletion of words, reordering or reorganizing sentences, and augmentation at the character level (e.g., simulating writing errors). These techniques have demonstrated their ability to improve results in text classification tasks, sentiment analysis, machine translation, and NLP-related tasks. Classification models benefit from better generalizing unseen data, dealing with overfitting, and other advantages.

In the case of track 3, when some of the described techniques were applied, the model showed a significant increase in its capacity, notably to avoid overgeneralization. In particular, the results obtained by creating new instances using word-embedding-based techniques with a BERT model showed notable results. This is significant considering that the result obtained in this task was the best presented among all teams, with data augmentation to address imbalance and overfitting being a fundamental factor. This technique was applied with null or more discrete results in the rest of the tasks.

4.3. Classification models

Transformer models have recently become very popular in Natural Language Processing (NLP). Thanks to their ability to capture complex contextual relationships in text, they have found a wide range

of applications in NLP, including hate speech detection. Transformers have established state-of-the-art in many NLP classification tasks, surpassing other machine learning models, so we chose to use different variants of this architecture to tackle the task at hand.

For each problem, we use different pre-trained models of architectures such as BERT, RoBERTa, and xlm-RoBERTa, specifically trained to deal with hate speech detection tasks on different platforms. We then adapt them to our specific dataset through additional training rounds, allowing better performance. Finally, we test the results of each model separately and variants of ensembles based on soft and hard voting techniques. For the soft-voting variants, we use simple heuristics to determine the weight of each model's vote, such as the average arithmetic sum. The results reveal that xlm-RoBERTa based models classify the proposed tasks more accurately. We decided to abandon the idea of ensemble architectures since, in general, the results of less accurate models worsened the result of the most accurate one unless their outputs were weighted as infinitesimal, which nullifies the advantages of using an ensemble.

5. Results and discussion

The results show that hate speech against the LGTB+ community can be identified and categorized with reasonable precision in the context of social media and cultural spaces in the Spanish language, specifically in the Mexican scenario. The sample size and the data quality used to create automated systems to address this task must be considered when deciding which methods to use to tackle the problem. A very effective method in one task with a given training set only sometimes generates good results in another. The adequate analysis of preprocessing techniques, data augmentation, and machine learning models significantly impacts the result obtained. However, in general, classifiers based on transformer architectures tend to generate superior results.

Table 8 shows the results for the three tracks. Our team obtained results comparable to those of the other teams and users. The most important result was first place in track three.

Table 8
Result for the three tracks in the test dataset with our procedure.

Track	F1-Score	Precision	Recall	Place
Track 1	85.59	90.15	82.57	6
Track 2	93.45	-	-	5
Track 3	57.62	56.03	65.12	1

6. Conclusion and future works

The study demonstrates the effectiveness of transformer-based models in detecting hate speech against the LGBTQ+ community in Spanish texts. The results show that transformer-based models, such as BERT, RoBERTa, and xlm-RoBERTa, outperform machine learning models to detect hate speech. The study highlights the importance of data augmentation in improving the performance of machine learning models in detecting hate speech. The results show that data augmentation techniques, such as word-embedding and synonym-based techniques, significantly improve the performance of the models. The research demonstrates the robustness of transformer-based models to overfitting in detecting hate speech. Transformer-based models are less prone to overfitting than machine-learning models. The Spanish language is a challenging domain for hate speech detection, the results show that detecting hate speech in Spanish texts is more challenging than other languages, such as English. In future works, we must prove another method for data augmentation and other classification models.

7. References

[1] Plaza-del-Arco, F. M., Molina-González, M. D., Urena-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, *166*, 114120.

[2] Chiruzzo, Luis and Jiménez-Zafra, Salud María and Rangel, Francisco (2024). Overview of IberLEF 2024: Overview of IberLEF 2024: Natural Language Processing Challenges. 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*.

[3] Vásquez, J., Andersen, S., Bel-Enguix, G., Gómez-Adorno, H., & Ojeda-Trueba, S. L. (2023, July). Homo-mex: A mexican spanish annotated corpus for lgbt+ phobia detection on twitter. In *The 7th Workshop on Online Abuse and Harms (WOAH)* (pp. 202-214).

[4] Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, *2*(6), 420.

[5] Guedes, G. B., & da Silva, A. E. A. (2024). Classification and Clustering of Sentence-Level Embeddings of Scientific Articles Generated by Contrastive Learning. *arXiv preprint arXiv:2404.00224*.

[6] Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139-172.

[7] Rothman, D. (2021). *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*. Packt Publishing Ltd.

[8] Arora, A., Mattoo, A., Chaudhary, D., Gorton, I., & Kumar, B. (2024, March). MEnTr@ LT-EDI-2024: Multilingual Ensemble of Transformer Models for Homophobia/Transphobia Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion* (pp. 259-264).

[9] Raj, V. S., Subalalitha, C. N., Sambath, L., Glavin, F., & Chakravarthi, B. R. (2024). ConBERT-RL: A policy-driven deep reinforcement learning based approach for detecting homophobia and transphobia in low-resource languages. *Natural Language Processing Journal*, *6*, 100040.

[10] Sharma, D., Gupta, V., & Singh, V. K. (2024). Abusive comment detection in Tamil using deep learning. In *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications* (pp. 207-226). Morgan Kaufmann.

[11] Khosla, C., & Saini, B. S. (2020, June). Enhancing performance of deep learning models with different data augmentation techniques: A survey. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 79-85). IEEE.

[12] Nagaraju, M., Chawla, P., & Kumar, N. (2022). Performance improvement of Deep Learning Models using image augmentation techniques. *Multimedia Tools and Applications*, *81*(7), 9177-9200.