Overview Homo-MEX 24 en IberLEF 2024: Hate speech detection towards the Mexican Spanish speaking LGBT+ population

Jorge Pérez Lanza[1,†], Hiram Ochoa[1,†], Jhovany Quintana[1,†] and Diego Iza[1,†]

[1] *Universidad Carlos III de Madrid, Av. de la Universidad, 30, 28911 Leganés, Madrid, Spain*

Abstract

Discrimination against the LGBT+ community continues to be a major global problem, manifested through hate speech, denial of rights and social exclusion. In social networks, platforms such as Twitter are scenarios where this type of discrimination prevails. The HOMO-MEX contest, which is part of the IberLEF 2024 initiative, focuses on the detection of hate speech targeting the Spanish-speaking LGBTQ+ community in Mexico. Our team applied advanced Natural Language Processing (NLP) techniques, using methods from Support Vector Machines (SVM) to Convolutional Neural Networks (CNN) with Short-Term Bidirectional Long Short-Term Memory (BiLSTM) networks and Transformer models, including the pre-trained Spanish BERT model and a variant of it in Mexican Spanish. The Transformer models consistently outperformed the other methods in identifying hate speech directed at the LGBT+ community. SVM models showed remarkable performance, while CNN-based models were less effective due to long training times. The importance of our study lies in the development of effective detection models and the creation of a specific corpus of Spanish-language LGBT+phobic annotated tweets, which fills a critical gap in existing resources and serves as a basis for future research.

Keywords

Hate Speech Detection, LGBT+ Discrimination, Mexican Spanish, Natural Language Processing, Support Vector Machines, Convolutional Neural Networks, Bidirectional Long Short-Term Memory, Transformer Models, BERT Spanish, Data Augmentation, Ensemble Learning, Social Media Analysis, IberLEF 2024, Corpus Development, Machine Learning Techniques

1. Introduction

Discrimination against the LGBT+ community remains a significant issue globally (Arimoro, 2022), manifesting in various forms such as hate speech, denial of rights, and social exclusion. In the context of social media, platforms like Twitter have become

[1*] Corresponding author.

[†] These authors contributed equally.

aleksandr.ometov@tuni.fi (A. Ometov); t.princesales@utwente.nl (T. P. Sales); manfred.jeusfeld@acm.org (M. Jeusfeld)

0000-0003-3412-1639 (A. Ometov); 0000-0002-5385-5761 (T. P. Sales); 0000-0002-9421-8566 (M. Jeusfeld)

arenas where such discrimination thrives (Vásquez et al., 2023). Recognizing the urgency to address this, the HOMO-MEX competition was established as part of the IberLEF 2024 initiative. This shared task focuses on the detection of hate speech directed at the Mexican LGBTQ+ community within the realm of Spanish, for this competition specifically Spanish Mexican.

Our team's contribution to this competition stems from a rigorous methodology combining advanced NLP techniques with a deep understanding of the linguistic nuances of Mexican Spanish. Leveraging methodologies ranging from Support Vector Machines (SVM) to Convolutional Neural Networks (CNN) paired with Bidirectional Long Short-Term Memory networks (BiLSTM) and Transformers, including the pre-trained BERT Spanish model and its Mexican variant, our study aimed to discern the most effective approach for detecting LGBT+Phobia in social media discourse.

Through meticulous experimentation and evaluation, we found that Transformer models consistently outperformed other methods, showcasing their efficacy in discerning hate speech targeting the LGBT+ community. Conversely, while SVM demonstrated notable performance, CNN-based models proved less effective, hindered by prolonged training times.

The significance of our work lies not only in the development of effective detection models but also in the creation of a dedicated corpus of Mexican Spanish tweets annotated for LGBT+Phobia. This corpus not only fills a critical gap in existing resources but also serves as a foundation for future research endeavors in this domain.

This article unfolds as follows: we begin by contextualizing the prevalence and ramifications of LGBT+Phobia in online discourse. Subsequently, we delve into the methodology employed, detailing the construction of our corpus and the experimental setup. We then present our findings, discussing the performance of various classification models and their implications. Finally, we conclude with reflections on the outcomes of our study and avenues for future research in combating LGBT+Phobia in digital spaces.

2. Related Work

2.1. Homomex 2023 iberLEF Competition

In the HOMO-MEX challenge presented at IberLEF 2023 (Bel-Enguix et al., 2023) developed a specific corpus for the detection of LGBTQ+phobic hate speech on Twitter in Mexican Spanish. The team implemented multi-label classification techniques, employing transformational models such as BERT and RoBERTa, and focusing on the importance of considering both context and implicit hate speech. Combined

approaches and preprocessing techniques were essential to improve accuracy in detecting and classifying LGBTQ+phobic content.

In the paper (Vásquez et al., 2023), several traditional machine learning algorithms and pre-trained deep learning models were used to establish a baseline classification of the corpus. The models evaluated included Support Vector Machines (SVM), Random Forests, and Transformer-based models such as BERT and its variants trained specifically for Mexican Spanish. The evaluation methodology was based on accuracy, recall, and F1-score to measure the performance of the models in the multi-label classification task. The results showed that deep learning models, in particular those based on Transformers, achieved better performance on the LGBT+ phobia detection task compared to traditional machine learning methods.

2.2. Classification Models for the Detection of Homophobia and Transphobia in Social Networks

In the paper (Kumaresan et al., 2023), traditional machine learning models and transformer-based deep learning models were used for homophobia and transphobia detection in social media comments in low-resourced languages. The models evaluated included pre-trained deep learning models, such as transformers, which showed improved performance compared to traditional methods. The evaluation of these models was performed using metrics such as accuracy, recall and F1-score. The results indicated that transformer-based models outperformed traditional methods in classifying homophobic and transphobic content, highlighting the effectiveness of deep learning models in moderating online content and promoting inclusive communities in social networks.

3. Dataset

The dataset for Track 1 consists of two columns: one containing tweets collected from Twitter, and the other containing the corresponding labels. According to the competition organizers, these categories are described as follows:

- 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 do not include any hate speech against the LGBT+ population but do mention this community.
- Not LGBT+related (NR): Tweets are not related in any way to the LGBT+ community.

Table 1. Some instances of the data and their corresponding classifications.

Content	Label
Me dan mucha risa los memes de la jota de caso cerrado 😂😂😂	P
#soyhomosensual #pepeyteo #marchalgbt #lgbt #gay #mexico #df @ Monumento a Cuauhtémoc url	NP
@LuisFrk eres un maricon culon	P
Olbaid666 @Cristiano Esta medio gay tu pregunta jajaja pero si el cabello largo viene de moda dale Micky	P
La jaula de las locas! (@ Teatro Hidalgo in Mexico City, Distrito Federal) url	NR
No sé que es peor los primeros 100 días de Trumpy o el puto tráfico de la ciudad. Creo el segundo, como q me causa mayor dolor de bolas.	NR

A preliminary exploration of the dataset revealed a total of 8800 records: 5482 labeled as NP, 2246 as NR, and 1072 as P. To train different models and perform hyperparameter optimization, the dataset was divided into three subsets: Train (75%), Validation (10%), and Test (15%). Stratified sampling was applied to ensure that the class proportions in the original dataset were maintained in each subset. Table 2 shows the distribution of the data after splitting.

Table 2. Division of data into Train, Val and Test Track 1.

Clase	Train Dataset	Validation Dataset	Test Dataset
P	820	91	161
NP	4194	466	822
NR	1718	191	1718

Track 2 focuses on the multi-class classification of tweets. This dataset comprises 1070 records and includes six possible categories or classes for classifying the tweets. The meanings of the labels are described below:

COLUMN LEGEND: L, G, B, T, O, NR

- Lesbophobia (L): Homophobia explicitly directed at homosexual people who identify as female.
- Gayphobia (G): Homophobia explicitly directed at homosexuals who identify as male.
- Biphobia (B): Hate speech directed against people who are attracted to more than one gender.
- Transphobia (T): Hate speech directed against non-cis-gendered people.
- Other LGBT+phobia (O): Hate speech against other sexual and gender minorities not included in the categories described above (e.g., "aphobia," which describes hatred towards people who do not feel sexual attraction).
- Not LGBT+related (NR): Tweets that are not related in any way to the LGBT+ community.

In summary, these datasets provide a robust foundation for evaluating different Machine Learning models' performance in identifying and classifying hate speech and other relevant categories within tweets. The careful stratification and detailed labeling ensure that models trained on these datasets will have a balanced and comprehensive understanding of the various types of content related to the LGBT+ community.

Table 3 shows a table with some instances and their corresponding classifications.

Table 3. Text instances corresponding to Track 2.

Content	L	G	B	T	O	NR
¡QUÉ POCA MADRE! ¡PERO EL JOTO @ManceraMiguelMX Y LA MACHORRA DE LA @CDHDF PREFIEREN DEFENDER A LOS MALDITOS DELINCUENTES! #PenaDeMuerteYa url	1	1	0	0	0	0
@JhonatanGaribay ajajajaja si no es que se huanguea tu puño y se vuelve puñal como manuela palma xD	0	1	0	0	0	0
Quieren un mundo #SinHomofobia pues que desaparezcan los jotos, maricones, putos, gays, lesbianas, machorras, tortilleras y demás sinónimos	1	1	0	0	1	0
Que mamada eh ! Seguramente todos los que irán a esa pendeja marcha, son mas closeteros que nada! #MarchaPorLaFamilia	0	1	0	0	1	0

In this case, the distribution of data associated with each class consists of 88 instances for class L, 894 for G, 10 for B, 94 for T, 77 for O, and 0 for NR. A simple analysis

suggests that the sum of all classes is greater than the number of cells in the dataset. This is due to the nature of the problem, as several tweets are assigned more than one class. For example, a review of the dataset shows that there are 20 instances of type L-G, meaning the tweet is classified as both "lesbophobic" and "gayphobic" simultaneously. Another important issue concerns the NR class, which has zero associated elements; the methodology section explains how the number of instances for this class was increased to 50. Similar to Track 1, the dataset for Track 2 was divided into three subsets. To achieve an equitable distribution, a new column was added to categorize the tweet based on the assigned classes. This led to various combinations and, therefore, new class distributions, which enabled stratification. It is necessary to clarify that, of the new classes created, instances with only one or two elements were excluded. Table 4 presents the division of the three datasets used in Track 2.

Table 4. Data Division in Train, Val, and Test for Track 2.

Clase	Train Dataset	Validation Dataset	Test Dataset
L	58	7	12
G	677	75	133
B	3	0	1
T	55	7	11
O	47	5	9
NR	39	4	7

A first conclusion derived from analyzing Table 4 is that class B is very poorly represented which can be a determinant in the performance of the models. How this problem was dealt with will be explained in another section.

4. Methodology

4.1. General approach

This section presents a summary of the methodology employed in this work. For this competition, ideas, approaches, and best practices were adopted from participants (García-Díaz et al., 2023) and (Moriña et al., 2023), who achieved good results in the Homomex 2023 competition. Similar to the approach described in (Vásquez et al.,

2023), in this work, several Machine Learning models were implemented to evaluate their performance in Tracks 1 and 2. Based on the obtained results, the model with the best performance was selected and established as the baseline. In other words, other developed models had to achieve better results than this pre-established baseline. This task was initiated with the hypothesis that Transformer models would have the best results, as they did last year [JP2]. Therefore, a starting point was established to determine the minimum expected performance.

An important step was the detailed review of the texts to create a cleaning function. A function was implemented to remove from the texts in each cell the presence of emojis, URLs, hashtags, mentions, and even additional white spaces. Eliminating these "noises" improves the estimation of future performance, and consequently, it can be stated that the models enhance their generalization capacity for both tasks.

Another crucial issue was the imbalance of the target classification classes for each task. In Track 1, it was observed that the NR and P classes were underrepresented, with approximately 25% and 12% respectively. This implies that a model trained with this dataset would tend to predict tweets from the NP class better than those from the NR and P classes. To address this situation, several solutions were proposed, including adjusting the "class_weight='balanced'" parameter in Machine Learning models from the Scikit-Learn library. The SMOTE (Synthetic Minority Over-sampling Technique) oversampling technique was also considered, as well as applying various transformations to the original data to create new versions, i.e., applying data augmentation. These techniques were selected for their potential benefits, although it was also considered that they could induce overfitting and generate very large datasets that could increase training time.

In Track 2, the imbalance is more critical than in Track 1. The G class has a representation of approximately 83% of the dataset, while the rest of the classes are all below 10%. Additionally, the NR class had no representation in the dataset, although it is considered one of the possible classification options. To address this problem, 50 NR-type instances extracted from the Track 1 dataset were incorporated into the dataset. This data injection was done before splitting the data for hyperparameter tuning and estimating the future performance of the different models.

In summary, the methodology employed in this work is based on the careful selection of Machine Learning models, thorough data cleaning, and the implementation of techniques to handle class imbalance, with the aim of improving the generalization capacity and performance of the models in both tracks of the competition.

4.2. SVM

Following the aforementioned steps, several Machine Learning models were evaluated, including RandomForest, DecisionTree, and SVM. For Track 1, as an initial step to process natural language data, the previously mentioned data cleaning was performed and then the data were transformed using CountVectorizer and TfidfTransformer.

CountVectorizer, a scikit-learn tool, converts a collection of text documents into a term-frequency matrix, creating a numerical representation of the text where each row represents a document and each column represents a term from the vocabulary; the values in the matrix indicate the frequency of each term's appearance in each document. The process involves tokenizing the text to split it into words or tokens, constructing a vocabulary with all unique words from the corpus, and counting the occurrences of each word in each document. Subsequently, TfidfTransformer converts this term-frequency matrix into a TF-IDF (Term Frequency-Inverse Document Frequency) matrix, a measure that reflects the importance of a term in a document relative to the entire corpus, reducing the weight of common terms and increasing that of rare terms, thus improving the model's ability to distinguish between relevant and irrelevant terms.

Among the analyzed models, the one that achieved the best results in terms of “macro avg f1-score” was the SVM in its Support Vector Classifier variant, with a macro avg f1-score of 0.73. However, this model obtained a recall of 0.36 for class P, which is considered low for the problem at hand due to the previously analyzed class imbalance.

To address this imbalance issue, the SMOTE oversampling technique was applied and the class_weight parameter was adjusted to 'balanced', but the results did not show a significant improvement. As a third option, data augmentation was implemented using a pre-trained model. A function was defined that loads a tokenizer and a pre-trained causal language model of RoBERTa. Tweets from the 'NR' and 'P' classes were filtered from the training set and randomly selected as input for generating new texts, ensuring that representative examples of these specific classes were used. Each selected tweet was tokenized and converted into a sequence of identifiers that the RoBERTa model can process. These identifiers were input into the model, which generated new texts with a maximum length of 500 tokens, using a temperature of 0.9 to introduce variability in the predictions. The generated texts were decoded back to their textual form and added to a list of generated texts, which were then transformed into a DataFrame.

It is important to note that, before this process, a stratified division of the dataset was performed, with 85% for training and 15% for testing. The training set was used as input for the data augmentation process, and the generated synthetic data were added to this set, ensuring no data leakage and that the model was not trained with instances

from the test set. Once the new training set was formed and the SVC model was trained with its corresponding hyperparameter search, the model's performance improved to a macro avg f1-score of 0.85, with a recall of 0.83 and 0.86 for the 'NR' and 'P' classes, respectively. With these performance values, it was determined that the SVC model prepared for Track 1 had an acceptable performance.

For Track 2, the methodology and functions used for data cleaning, transformations, and vectorizations were the same as those employed in Track 1. Due to the multiclass nature of this task, the columns containing the classes were processed. A function was developed that for each row adds the corresponding letters to the columns where the value is 1 and then concatenates them into a new column called 'label'. This resulted in the creation of several new classes, such as 'GO', 'LG', 'TO', 'GT', 'GB', 'LGTO', among others. Based on the frequencies of these classes, it was decided to eliminate those with only one or two instances within the dataset.

This transformation allowed for a stratified division of the training and test sets. To address the imbalance issue, 50 instances of the 'NR' class were injected into the dataset since this class had zero rows. The 'B' class was also poorly represented, so it was decided to apply data augmentation directly. Although the same process as in Track 1 was attempted, it was not satisfactory due to the small number of examples associated with this class. In this case, what worked best was using a prompt model, specifically GPT-3.5, to generate 50 examples with similar length and semantics to those of the 'B' class. Once the splits for the training and evaluation sets were made, the newly created column was discarded and the synthetic data were added to the training set to improve the imbalance.

Regarding the SVC, the approach that yielded the best results was using the "One-vs-Rest" (OvR) classifier. This resulted in a 0.89 in the “micro avg f1-score” metric. Although the recall for some classes was not optimal, it was determined that the estimated future performance of this model was adequate, and thus it was used for the competition.

4.3. Transformers

Track 1

In the context of Track 1, two transformer models, BETO and BETO_Clasificar_Tweets_Mexicano, were employed to classify tweets related to LGBT+phobia. Data distribution was the same as in SVM. Both BETO-based models were evaluated under conditions with and without data augmentation. Initially, default hyperparameters were used, followed by hyperparameter optimization, resulting in the final models. These BETO-based models showed strong performance, reflecting the potential of transformer models in effectively identifying hate speech in Mexican Spanish tweets.

The BETO model was fine-tuned on a dataset for detecting LGBT+phobic content in tweets, specifically using data from the HomoMex 2024 competition. For tokenization, the appropriate tokenizer for the BETO model was loaded, and the maximum token length in the training dataset was determined. The pre-trained BETO model was then fine-tuned for sequence classification. Training arguments, including batch sizes and learning rate, were defined, and a Trainer object from the transformers library was created to handle the training and evaluation processes.

Table 5. Hyperparameters selected for transformers in track 1.

Hyperparameter	Value
per_device_train_batch_size	32
per_device_eval_batch_size	32
learning_rate	5e-5

During training and evaluation, the model was trained using the training dataset and evaluated on the validation dataset. Predictions were generated for the test dataset, and a classification report was computed, including precision, recall, and F1 score for each class. A confusion matrix was plotted to visualize the model's performance.

Track 2

In the context of Track 2, the same models used in Track 1 were employed. Unlike the previous task, which focused on single-label classification, this approach addresses a multi-label classification problem, allowing a tweet to be categorized into multiple classes simultaneously, reflecting the complexity and overlapping nature of hate speech content.

The BETO-based models were fine-tuned for a multi-label classification problem. The tokenization process followed a similar procedure as in Track 1, with the appropriate tokenizer loaded and the maximum token length determined. The pre-trained BETO model was fine-tuned, tailored for a multi-label classification problem. Training arguments were defined, and a Trainer object was created to handle the training and evaluation processes.

Table 6. Hyperparameters selected for transformers in track 2.

Hyperparameter	Value
per_device_train_batch_size	32

During training and evaluation, the model was trained using the training dataset and evaluated on the validation dataset. Predictions were generated for the test dataset, and a classification report was computed, including precision, recall, and F1 score for each class. A confusion matrix was plotted to visualize the model's performance.

4.4. CNN + BiLSTM

The CNN layer transforms data to perform linear and nonlinear functions. The CNN layer has a two-dimensional kernel that converts the input sequence into two dimensions. A CNN neural network consists of numerous two-dimensional arrays of neurons. Each neuron is connected to its small neighboring area of neurons in the previous layer through a feed-forward connection. The activation features of the Rectified Linear Unit (ReLU) introduce non-linearity to the neural data.

(colocar la cita del autor unas dos)

To implement the dataset with these models, we applied various techniques based on the modified dataset function. It is worth mentioning that the CNN+BiLSTM models are designed for learning with images, but they can also be implemented with text-based datasets.

Track1

To implement the model, we first examined the distribution of the dataset using three-label classification. We found that data augmentation was necessary to improve the quality of training and obtain more accurate results. To enhance the outcomes, we initially employed word embedding techniques. However, the performance was very poor in terms of the F1 metric.

We attempted to improve the model by adding a new layer to the CNN and modifying the hyperparameters. Despite these efforts, the results were not satisfactory. The inclusion of word embeddings did not enhance the model's performance as expected.

Additionally, we tried incorporating several convolutional network layers into the model. However, this resulted in even lower performance, indicating that the additional complexity did not benefit the model in this situation.

Finally, we used the data augmentation technique. Since it significantly improved the model's performance, this strategy proved to be much more effective. Data augmentation allowed us to create a more diverse and representative dataset, which improved the model's ability to generalize and recognize patterns in the input data. This strategy enabled us to achieve much more optimal and satisfactory results.

Track 2

Each of the five records in the dataset has a binary result, with a value depending on the corresponding label. Since the initial results were regular, a data augmentation technique was not used, even though the dataset is relatively small. Although we used different techniques such as word embedding, the results were not satisfactory.

We implemented the same CNN+BiLSTM layer structure, but the results were significantly lower because the dataset structure contained more labels. We tried adjusting the model by removing some of the default CNN layers, but the results in terms of binary classification remained unsatisfactory.

Additionally, we added an extra layer to the model to evaluate its impact, but the results continued to be very low. This experiment indicated that the additional complexity and layers did not benefit the model's performance with the specific dataset we were using. Binary classification, in this case, did not improve with the removal or addition of layers in the CNN+BiLSTM structure. The best results were obtained by using standard layers for optimal performance.

5. Results and Discussion

5.1.1. SVM

Once the simulations and the corresponding hyperparameter searches were executed, a Support Vector Classifier model was obtained with an F1-score of around 0.85, with the class 'NP' having the lowest recall. Under these performance conditions, the model's results were deemed acceptable. The competition results were an F1-score of 0.7818236078707702.

For the track, a performance of 0.89 in micro avg F1-score was estimated. However, to the authors' surprise, the official results were 0.9322139303, which positioned this model very well.

5.1.2. Transformers

Track 1

Original dataset

The table below presents the results of the transformer models evaluated on the original dataset for Track 1. Both models, BETO (bert-base-spanish-wwm-uncased) and BETO_Clasificar_Tweets_Mexicano, were assessed using key performance metrics: accuracy, precision, recall, and F1-score. The F1-score, in particular, is critical as it balances precision and recall, providing a single metric that reflects the model's overall performance. As shown in the table, both models achieved identical performance metrics, indicating that their ability to classify LGBT+phobic content in tweets is highly comparable. Both models exhibit a high level of accuracy and balanced precision and recall, underscoring their robustness in this application.

Table 7. Track 1 results transformers original dataset.

Model	accuracy	precision	recall	F1-score
bert-base-spanish-wwm-uncased	0.87	0.83	0.81	0.82
BETO_Clasificar_Tweets_Mexicano	0.87	0.83	0.81	0.82

Data Augmentation

The table below presents the results of the transformer models evaluated on the augmented dataset for Track 1. Data augmentation techniques were applied to enhance the dataset, potentially improving model performance by increasing the diversity of training examples. Both models, BETO (bert-base-spanish-wwm-uncased) and BETO_Clasificar_Tweets_Mexicano, were assessed using the same performance metrics: accuracy, precision, recall, and F1-score. As shown in the table, data augmentation significantly improved the performance metrics for both models.

Table 8. Track 1 results transformers with data augmentation.

Model	accuracy	precision	recall	F1-score
bert-base-spanish-wwm-uncased	0.93	0.89	0.94	0.91
BETO_Clasificar_Tweets_Mexicano	0.90	0.94	0.92	0.92

These results indicate that the introduction of data augmentation positively impacted the performance of the models. The BETO_Clasificar_Tweets_Mexicano model, in

particular, demonstrated a higher F1-score compared to the original dataset results, suggesting enhanced robustness and better generalization capabilities when exposed to a more diverse set of training examples. The overall improvement in accuracy, precision, recall, and F1-score underscores the effectiveness of data augmentation in enhancing the models' ability to classify LGBT+phobic content in Mexican Spanish tweets.

Since the BETO-based models achieved the best results among all the models tested in this paper, the confusion matrix for the BETO model will be presented.

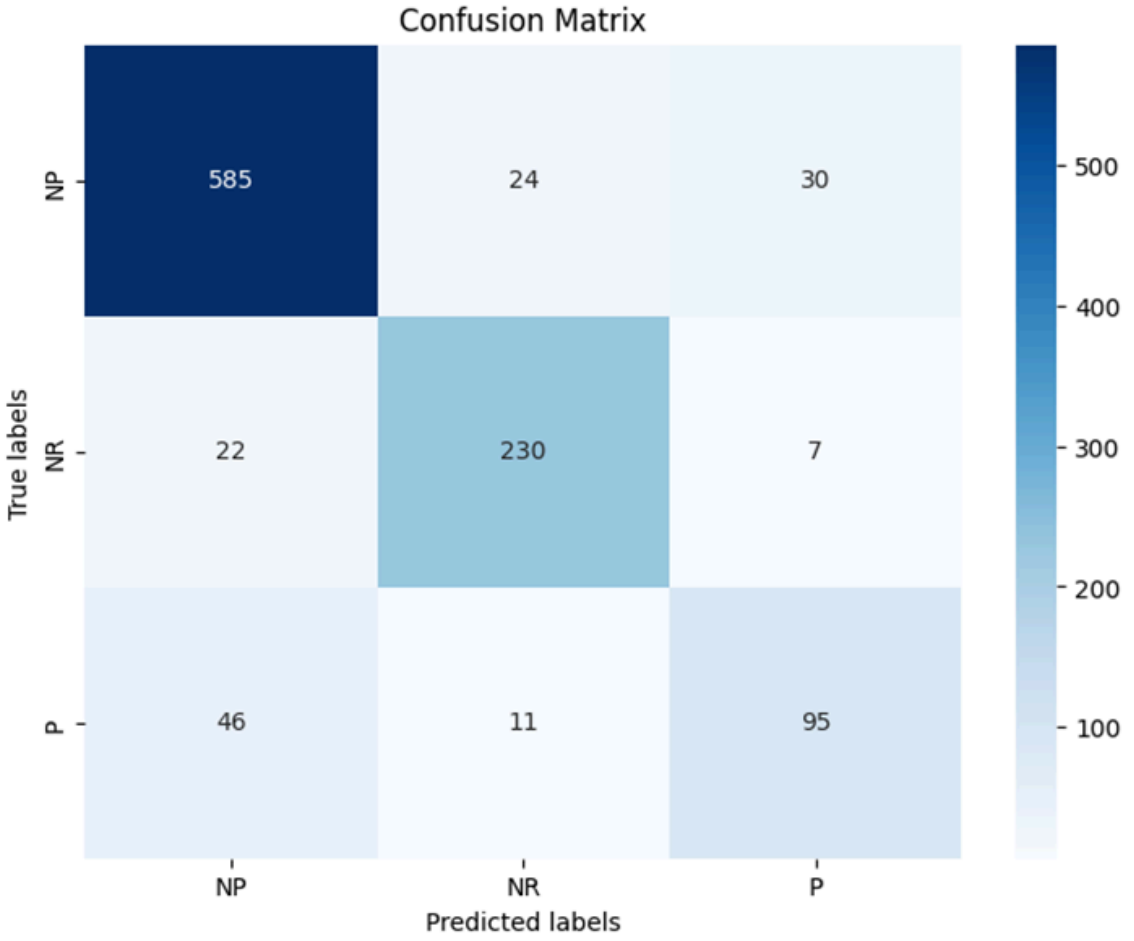

Figure 1. Confusion matrix of BETO without data augmentation.

According to the figure above, the BETO model without data augmentation performs well in correctly identifying the NP class, with 585 correct predictions and relatively few misclassifications (24 as NR and 30 as P). For the NR class, the model shows strong performance with 230 correct predictions, but with higher misclassifications as NP (22 instances) compared to the results with data augmentation. The P class has 95 correct

predictions, with a higher number of misclassifications as NP (46 instances) and 11 instances as NR, compared to the augmented dataset

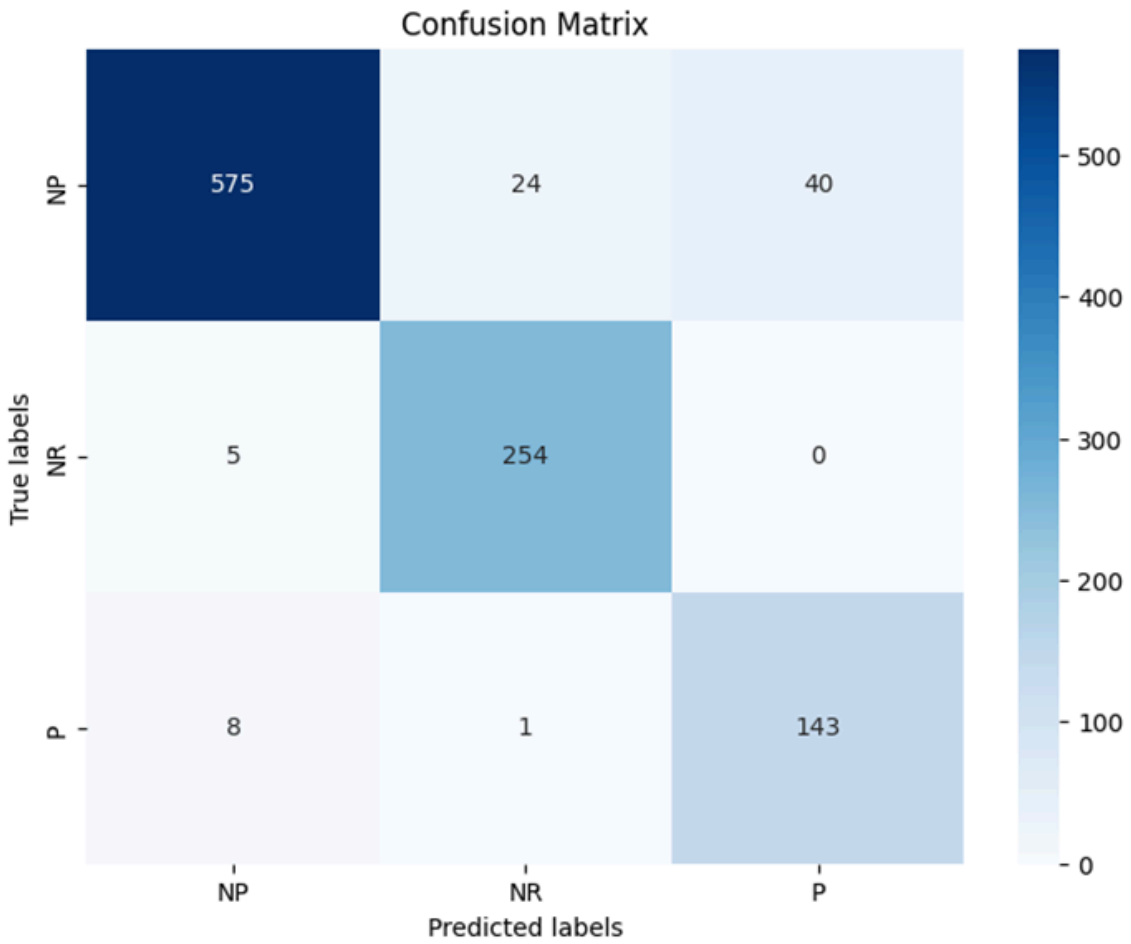

Figure 1. Confusion matrix of BETO with data augmentation.

According to the figure above, the BETO model with data augmentation performs well in correctly identifying the NP class, with 575 correct predictions and relatively few misclassifications (24 as NR and 40 as P). For the NR class, the model also shows strong performance with 254 correct predictions and only 5 misclassifications as NP. The P class has 143 correct predictions, with 8 instances incorrectly predicted as NP and only 1 instance as NR.

Track 2

Original dataset

The table below presents the results of the transformer model evaluated on the original dataset for Track 2. This track focuses on multi-label classification, allowing tweets to be categorized into multiple classes simultaneously. The model assessed is BETO

(bert-base-spanish-wwm-uncased), and the performance metrics used are precision, recall, and F1-score. As shown in the table, the BETO model achieved high precision, recall, and F1-score, demonstrating its effectiveness in handling the complexity of multi-label classification for detecting LGBT+phobic content in tweets.

Table 9. Track 2 results transformers original dataset.

Model	precision	recall	F1-score
bert-base-spanish-wwm-uncased	0.94	0.88	0.91

These results highlight the strong performance of the BETO model in accurately identifying and classifying various types of LGBT+phobia in Mexican Spanish tweets. The high precision and recall values indicate that the model is capable of correctly identifying relevant instances while minimizing false positives and false negatives, respectively. The F1-score further reflects the model's overall balanced performance in this multi-label classification task.

5.1.3. CNN + BiLSTM

"The precision and F1 scores of CNN+BiLSTM are higher. This indicates that, although it may not find all the positives (lower recall), the predictions it makes are more precise and it has a better balance between precision and recall.

The precision and recall improve with CNN+BiLSTM+Word embedding. This indicates that finding most of the actual positives is better for the model, but there may be more false positives, which reduces its precision and F1 scores."

Track 1

Table 10. Track 1 results transformers original dataset.

Model	accuracy	precision	recall	F1-score
CNN+BiLSTM	0.40	0.46	0.42	0.78

CNN+BiLSTM+Word embedding	0.60	0.42	0.82	0.43

"The CNN+BiLSTM model outperforms the CNN+BiLSTM+Word Embedding model in all metrics. This indicates that the combination of CNN+BiLSTM without word embeddings is more effective and balanced."

Track 2

Table 11. Track 2 results transformers original dataset.

Model	accuracy	precision	recall	F1-score
CNN+BiLSTM	0.33	0.35	0.32	0.49
CNN+BiLSTM+Word embedding	0.23	0.15	0.21	0.35

6. Official Results

The participant models had the following local metrics before being evaluated by HomoMex staff.

Table 12. Track 1 local results comparison.

Model	accuracy	precision	recall	F1-score
bert-base-spanish-wwm-uncased	0.93	0.89	0.94	0.91
SVM	0.87	0.82	0.88	0.85

CNN+BiLSTM	0.40	0.46	0.42	0.78

Table 13. Track 2 local results comparison.

Model	precision	recall	F1-score
bert-base-spanish-wwm-uncased	0.94	0.88	0.91
SVM	0.93	0.84	0.89
CNN+BiLSTM	0.15	0.21	0.35

The official results of the competition are presented below. We participated with the username "jlpl1" and achieved the following rankings, both in track 1 and track 2:

Table 14. Track 1 official rankings.

Rank	Username	F1-Score	Precision	Recall
1	verbanex	91.43	93.64	89.63
2	atoro491	91.43	93.64	89.63
3	rogerd97	91.43	93.64	89.63
9	**jlpl1**	**84.18**	**90.64**	**78.57**

Table 15. Track 2 official rankings.

Rank	Username	F1-Score	Hamming Loss	Exact Match Ratio
1	quanle709	97.30	14.93	92.91
2	homomex	94.88	23.63	88.43
3	sdamians	94.35	34.20	84.70
7	**jlpl1**	**93.22**	**27.98**	**89.93**

Summary:

- In Track 1, we secured the 9th place with an F1-Score of 84.18.
- In Track 2, we achieved the 7th place with an F1-Score of 93.22.

7. Conclusion

In this paper, we have presented our approach and findings in the HOMO-MEX competition, which focused on detecting hate speech towards the Mexican Spanish-speaking LGBTQ+ community. Our participation included the application of advanced natural language processing techniques, specifically leveraging transformer models such as BERT and its variations, to tackle the challenges presented in both tracks of the competition.

For Track 1, our methodology involved the use of data augmentation and hyperparameter optimization to improve model performance. The results demonstrated the efficacy of these approaches, with our best model achieving an F1-score of 84.18, placing us in 9th position. This indicates that while our methods were effective, there remains potential for further improvement, particularly in addressing the class imbalance present in the dataset.

In Track 2, we faced the additional complexity of multi-label classification. Here, we employed a combination of transformer models with ensemble learning techniques and SVM with data augmentation. Our efforts resulted in a commendable F1-score of 93.22 for the SVM model, securing the 7th position. This highlights the robustness of our

models in handling complex classification tasks involving multiple overlapping categories.

Overall, our participation in the HOMO-MEX competition has underscored the importance of sophisticated machine learning techniques in detecting and classifying hate speech. The use of transformers, data augmentation, and ensemble methods proved particularly beneficial. Moving forward, future work will focus on further enhancing model performance through additional data preprocessing techniques, exploring more advanced ensemble strategies, and addressing the inherent class imbalances in the datasets. Additionally, expanding the corpus with more diverse examples of hate speech can help improve the generalizability and accuracy of the models.

Our findings contribute to the broader goal of developing effective tools for moderating online content and promoting a safer and more inclusive digital environment for the LGBTQ+ community. Through continuous innovation and collaboration, we aim to advance the field of hate speech detection and support the ongoing fight against online discrimination.

8. Future Work

Our study has laid a solid foundation for detecting and classifying hate speech directed at the LGBTQ+ community within Mexican Spanish tweets. However, there are several avenues for future work that could enhance the efficacy and robustness of our models and methodologies.

Firstly, addressing the class imbalance remains a critical challenge. While data augmentation techniques were employed, further exploration into advanced methods such as Generative Adversarial Networks (GANs) for synthetic data generation could provide more balanced datasets and improve model training.

Secondly, expanding the dataset to include a broader range of social media platforms beyond Twitter could help create a more comprehensive corpus. Platforms like Facebook, Instagram, and TikTok have different user demographics and communication styles, which can introduce additional linguistic variations and contexts of hate speech.

Thirdly, integrating additional contextual information into the models could enhance performance. This includes metadata such as the time of posting, user demographics,

and geolocation data, which can provide richer context and help the models understand the nuances of hate speech better.

Moreover, exploring ensemble methods further and combining them with other advanced machine learning techniques, such as transfer learning from larger, more diverse pre-trained models, can potentially yield better results. This approach can leverage the strengths of various models and reduce the weaknesses of individual classifiers.

Another promising area is the implementation of explainable AI techniques to make the model predictions more transparent. Understanding why a model classifies a particular tweet as hate speech can help in refining the models and making them more reliable and acceptable to stakeholders.

Finally, ongoing collaboration with linguists, social scientists, and the LGBTQ+ community is essential. Their insights can guide the development of more culturally and contextually relevant models, ensuring that the technology remains aligned with the needs and sensitivities of the communities it aims to protect.

References

Bel-Enguix, G., Gómez-Adorno, H., Sierra, G., Vásquez, J., Andersen, S. T., & Ojeda-Trueba, S. (2023). 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*, 361–370. https://doi.org/10.26342/2023-71-28

Kumaresan, P. K., Ponnusamy, R., Priyadharshini, R., Buitelaar, P., & Chakravarthi, B. R. (2023). Homophobia and transphobia detection for low-resourced languages in social media comments. *Natural Language Processing Journal*, *5*, 100041. https://doi.org/10.1016/J.NLP.2023.100041

Vásquez, J., Andersen, S. T., Bel-Enguix, G., Ojeda-Trueba, S.-L., & Gómez-Adorno, H. (2023). *HOMO-MEX: A Mexican Spanish Annotated Corpus for LGBT+phobia Detection on Twitter*.

Arimoro, A. E. (Ed.). (2022). Global perspectives on the LGBT community and non-discrimination. IGI Global.

García-Díaz, J. A., Jiménez-Zafra, S. M., & Valencia-García, R. (2023). UMUTeam at HOMO-MEX 2023: Fine-tuning Large Language Models integration for solving hate-speech detection in Mexican Spanish.

Moriña, A. J. M., Pásaro, J. R., Vázquez, J. M., & Álvarez, V. P. (2023). I2C-UHU at IberLEF-2023 HOMO-MEX task: Ensembling Transformers Models to Identify and Classify Hate Messages Towards the Community LGBTQ.

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