I2C-UHU at IberLEF-2023 HOMO-MEX task: Ensembling Transformers Models to Identify and Classify Hate Messages Towards the Community LGBTQ+

Antonio José Morano Moriña, Javier Román Pásaro, Jacinto Mata Vázquez and Victoria Pachón Álvarez

I2C Research Group ,University of Huelva, Spain

Abstract

This paper presents the approaches proposed for I2C Group to address the IberLef-2023 Task HOMO-MEX: Hate speech detection in Online Messages directed tOwards the MEXican spanish speaking LGBTQ+ population. The major contribution has been the demonstration of the effectiveness of using an ensemble of classifiers based on transformers. By combining multiple models, the individual strengths were leveraged, resulting in improved performance compared to using a single model. Furthermore, the significance of selecting appropriate hyperparameters during the model training process was underscored by the results. Through meticulous experimentation and evaluation of different hyperparameter combinations, the settings that reached the best performance for the given tasks were identified. In our experiments for both tasks we have tested several models and decided to ensemble the three models that provided the best F1-Score for this dataset. Additionally, for Task 2 we decided to train individual binary classifiers for each class instead of making a multilabel classifier. The model submitted for Task 1 achieved a F1-Score of 83,25%, ranking in the 6th place of the competition. The model for the Task 2 reached a F1-Score of 69,60%, ranking in the 1st place of the competition.

Keywords

Deep Learning, Transformers, Ensembler, Hyperparameter, Twitter, LGBT-Phobia, Hate Speech Detection

1. Introduction

In today's digital era, natural language processing (NLP) has become an essential discipline for understanding and analyzing the vast amount of information generated on social media platforms. The ability to extract meaningful knowledge from textual data is crucial for various fields, including social research, political decision-making, and the detection of social issues. In this context, the detection of phobic comments towards the LGBTQ+ community has gained increasing importance due to the need to promote inclusion, respect, and equality online.

This paper presents our research on developing a system for detecting phobic comments towards the LGBTQ+ community using natural language processing techniques as part of the HOMO-MEX: Hate speech detection towards the Mexican Spanish speaking LGBTQ+ population

IberLEF 2023, September 2023, Jaén, Spain

antoniojose.morano490@alu.uhu.es (A. J. M. Moriña); javier.roman780@alu.uhu.es (J. R. Pásaro); mata@uhu.es (J. M. Vázquez); vpachon@dti.uhu.es (V. P. Álvarez)

CEUR Workshop Proceedings (CEUR-WS.org)

from IberLEF 2023 [1] [2] task. Given the success and popularity of Transformers models [3], all the developed models are based on this technology. In order to get our final results, we trained three models and built an ensemble [4] to improve classifier performance in both tasks. Additionally, for the multilabel task, we decided to use individual binary classifiers instead a multilabel classifier.

In the next section some previous studies are described. In Section 3 we will describe Tasks 1 and 2 and the Corpus provided by the organizers. The experimental methodology and evaluation results can be found in Section 4 and 5. Finally, in Section 6, the conclusions of our study are shown and some perspectives for future works are described.

2. Related works

In recent years, the field of Natural Language Processing (NLP) has witnessed significant advancements, particularly with the advent of transformer models. These models, such as BERT [5], GPT [6], and RoBERTa [7], have revolutionized the way we process and understand text, enabling us to tackle complex linguistic tasks with unprecedented accuracy. One crucial application of NLP technology is the detection of hate messages and discriminatory content, particularly those targeting marginalized communities like the LGBTQ+ community.

Several recent investigations have focused on leveraging transformer models for detecting hate messages against the LGBTQ+ community. For example, [8] explored the use of pre-trained transformer models for hate speech detection and found that fine-tuning these models on annotated LGBTQ+ hate speech datasets significantly improved their performance . By leveraging the contextualized representations learned by transformer models, the researchers were able to capture the subtle nuances and linguistic patterns indicative of hate speech.

These recent investigations showcase the potential of transformer-based models in detecting hate messages against the LGBTQ+ community. By training these models on large, annotated datasets and fine-tuning them specifically for hate speech detection, researchers have achieved significant advancements in accurately identifying and categorizing discriminatory content. 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 [9].

3. Datasets and Tasks

The Corpus provided by the organizers is described at Codalab (https://codalab.lisn.upsaclay.fr/competitions/10019). This Corpus contains two datasets, one per task:

- The first one consists of 7000 tweets formed by an identifier, the tweet text, and the label of the instance. Three different labels were defined: LGBT+phobic (P), not LGBT+phobic (NP) or not LGBT+related (NA). Since the organizers provided only one dataset, we decided to divide it into training (80%), validation (14%), and test (6%).

Table 1
Class distribution for Task 1

Class	Train Dataset	Valid Dataset	Test Dataset
P	690	249	107
NP	3488	610	262
NA	1422	121	51
Total	5600	980	420

Table 2
Some instances of Task 1

Index	Tweet	Label
92	Nada más peligroso que un joto con autoestima demasiado alto! (Nothing more dangerous than a gay man with excessively high self-esteem!)	P
2237	@marisita_parra entonces ser homosexual es no tener valores? No sé de que hablas. (@marisita_parra is being homosexual synonymous with having no values? I don't know what you're talking about.)	NP
441	Esta noche es perfecta para volverte loca (Tonight is perfect to drive you crazy.)	NA

- For the second task, the dataset contains 863 tweets, with the same information and five different labels: Lesbophobia (L), Gayphobia (G), Biphobia (B), Transphobia (T), and/or other LGBT+phobia (O).

Table 3
Class distribution for Task 2

Label	Train Dataset	Valid Dataset	Test Dataset
G - Phobic	575	99	40
G - Non Phobic	114	22	12
L - Phobic	57	9	6
L - Non Phobic	632	112	46
B - Phobic	8	1	1
B - Non Phobic	681	120	51
T - Phobic	57	16	6
T - Non Phobic	632	105	46
O - Phobic	48	12	4
O - Non Phobic	641	109	48

Tables 1 and 3 show the distribution of the classes for each task, after the split into training, validation, and test. Tables 2 y 4 show some examples of the tweets that the datasets respective to the Task 1 and 2 contain.

Table 4
Some instances of Task 2

Tweet	G	L	B	T	O
Quieren un mundo #SinHomofobia pues que desaparezcan los jotos, maricones, putos, gays, lesbianas, machorras, tortilleras y demás sinónimos (They want a world #WithoutHomophobia so let the homosexuals, fags, hustlers, gays, lesbians, butchers, dykes and other synonyms disappear.)	1	1	0	0	1
Me reeemputa que dejen jugar mujeres trans en torneos femeniles, como vergas bloqueas a un cabron de 1.80 que pesa el doble que tú y tiene el triple de fuerza (It pisses me off that they let trans women play in women's tournaments, how the fuck do you block a 6'4" motherfucker who weighs twice as much as you and is three times as strong?)	0	0	0	1	0
¿Cómo qué hay mujeres trans lesbianas? ¿Para que se hizo trans si va a ser lesbiana? No tiene lógica. (Why are there lesbian trans women? Why did she become trans if she's going to be a lesbian? It doesn't make sense.)	0	1	0	1	0

4. Methodology

This section outlines the methodology employed in this study, which consisted of several key steps. Firstly, due to the lack of data in the phobic class, a data augmentation approach based on the backtranslation technique was used. Secondly, a hyperparameter search was conducted to identify the optimal training parameters for this particular task. Finally, a clasification model was created by ensembling the three best found models and implementing a hard voting approach in order to enhance performance.

Because the datasets are in Spanish language, pre-trained Spanish models were used primarily. However, given that Mexican Latin American Spanish contains a significant amount of Anglo-Saxon vocabulary, a multilingual model was also chosen to explore alternative options. The pre-trained models selected, obtained from the Hugging Face Transformers library (https://huggingface.co/), were:

- dccuchile/bert-base-spanish-wwm-uncased [10]. This model (BETO) is a BERT Spanish version
- PlanTL-GOB-ES/roberta-base-bne [11]. This model is based on the RoBERTa base model and has been pre-trained using the largest Spanish corpus known to date
- xlm-roberta-base [12]. This model is a multilingual version of RoBERTa.

To compare the results obtained by the different models and developed strategies, a baseline based on the pre-trained selected models was proposed. Given that it is not possible to know the optimal values of the hyperparameters beforehand, some of the most frequently used values were employed to perform fine-tuning of pretrained language models: batch size of 32, learning rate of 5e-5, max length of 128 and weight decay of 0.001. Tables 5 and 6 show the baseline results on different models for tasks 1 and 2.

Table 5
Baseline results for Task 1

Model	F1-Score (Macro Average)
BETO	0.8172
RoBERTa	0.8011
XLM	0.8227

Table 6
Baseline results for Task 2

Model	F1-Score (Macro Average)
BETO	0.6412
RoBERTa	0.6318
XLM	0.6128

4.1. Data Pre-processing

The data pre-processing consisted on removing links, usernames, hashtag symbols '#', and emojis. Additionally, we created a dictionary of synonyms (https://es.wiktionary.org/wiki/Wikcionario:homosexual/Tesauro) where words specific to Mexican Spanish language were replaced with more common alternatives that had the same meaning but fit into the vocabulary of the pre-trained models. Tables 7 and 8 show the results achieved after processing the texts from the tweets. As it can be seen, this pre-proccessing improved the results obtained with the baselines.

Table 7
Results with Pre-processing for Task 1

Model	F1-Score (Macro Average)
BETO	0.8281
RoBERTa	0.8197
XLM	0.8228

Table 8
Results with Pre-processing for Task 2

Model	F1-Score (Macro Average)
BETO	0.6503
RoBERTa	0.6726
XLM	0.6539

4.2. Data Augmentation and Hyperparameter Search

In order to balance the multiclass dataset, a data augmentation based on a backtranslation technique [13] was used. This technique was applied to increase instances of the class P (Phobic) in the dataset, doubling the number of phobic instances. For multilabel dataset, backtranslation technique was applied to the complete dataset, increasing the positive instances of each label in a 50%. A translation from Spanish to English and backwards was carried out. The pre-trained model "Helsinki-NLP/opus-mt-es-en" [14] was utilized for the first translation, and the model "Helsinki-NLP/opus-mt-en-es" [15] was used for the backtranslation.

The hyperparameter search [16] is a crucial step for models fine-tuning. For this reason, multiple iterations of training and evaluation were performed using different combinations of some hyperparameters. To reduce training time costs, the datasets were proportionally reduced before conducting the experimentation. The platform used for this purpose was WandB (Weights & Biases, wandb.com) , which provides a clear graphical interface for tracking and visualizing machine learning experiments. Table 9 shows the hyperparameter space used in this experimentation phase.

Table 9
Hyperparameters space

Hyperparameter	Values
Batch Size	[16, 32, 64]
Learning Rate	[2e-5, 3e-5, 5e-5]
Max Length	[64, 128, 256]
Weight Decay	[0.001, 0.01, 0.1]

In Table 10 we can see the best hyperparameters found for each model. Tables 11 and 12 show the results of each model using data augmentation and the hyperparameters values from Table 10. The results showed in Tables 11 and 12 prove the importance of working with a balanced dataset and performing a proper hyperparameter search for an optimal fine-tuning.

Table 10
Best Hyperparameters per model

Hyperparameter	BETO	RoBERTa	XLM
Batch Size	32	32	16
Learning Rate	5e-5	3e-5	2e-5
Max Length	128	128	256
Weight Decay	0.01	0.01	0.01

4.3. Ensemble Approach

To make the final predictions, a hard voting technique [17] was implemented. The most common prediction among the models was chosen as the final output, ensuring a more robust and consensus-based prediction. The ensemble [18] and model voting techniques helped enhance

Table 11
Results with Data Augmentation and Hyperparameter Search for Task 1

Model	F1-Score (Macro Average)
BETO	0.8566
RoBERTa	0.8451
XLM	0.8228

Table 12
Results with Data Augmentation and Hyperparameter Search for Task 2

Model	F1-Score (Macro Average)
BETO	0.6674
RoBERTa	0.6960
XLM	0.6714

the overall predictive performance by leveraging the strengths and diversity of multiple models, leading to more accurate and reliable predictions. For both tasks, the models used in the ensembles were the ones described earlier, namely BETO, RoBERTa, and XLM. In the event that the three individual predictions differ, the selection of the final prediction would prioritize the model with the highest F1-Score.

For the multi-label classifier, an ensemble approach was implemented on a per-label basis. This involved creating separate ensembles for each label by concatenating the results of the five individual predictions specific to that label. By combining these predictions, a final output for each label was generated.

These results reflect the collective decision-making of the models and represent the final outcome that were uploaded for assessment in the competition.

5. Results

In this section, we present the final results submitted for the two tasks. The predictions were evaluated using the official competition metrics, specifically the macro F1-Score.

For Task 1, the final prediction was constructed using a voting scheme among the three models, with BETO acting as the tiebreaker. The achieved F1-Score for this task was 0.8325, resulting in a sixth position. Table 13 shows the final leaderboard for Task 1.

For Task 2, RoBERTa had the ability to determine the outcome in the event of a tie between the three models because it is the model with the highest F1-Score. This results in a F1-Score of 0.6960 obtaining the first place in the competition. Table 14 shows the final leaderboard for Task 2.

The obtained rankings demonstrate the effectiveness of our approach and the promising outcomes achieved.

Table 13
Ranking of participants for Task 1

Ranking	User	Prediction F1-Score
1	bayesiano98	0.8847
2	carfer	0.8432
3	JoseAGD	0.8421
4	homomex23	0.8390
5	Cordyceps	0.8354
6	**I2C - Huelva**	**0.8325**
-	-	-
11	moeintash	0.7326

Table 14
Ranking of participants for Task 2

Ranking	User	Prediction F1-Score
1	**I2C - Huelva**	**0.6960**
2	carfer	0.6847
3	ErikaRivadeneira	0.6834
-	-	-
9	cesar_m	0.6550

6. Error Analysis

The confusion matrices of the classifiers for both tasks on our test dataset can be found in Figure 1 and 2.

Figure 1 shows how well the classifier performs when predicting classes NP (Not Phobic) and (Not Related) in the Task 1. Even so, it is not as reliable at predicting class P (Phobic). This may be the result of the large imbalance in the training dataset, where the phobic class has the lowest presence.

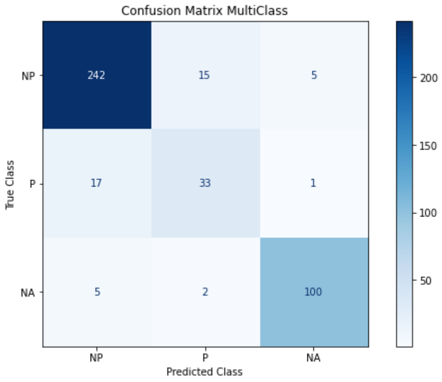

Figure 1: Confusion Matrix for Task 1

Although good results have been obtained in Task 1, it must be borne in mind that on rare occasions errors have been made in the prediction. Table 15 shows some of the few instances where errors have been made. The limited characters and lack of context with such similar vocabulary can lead to confusion.

Table 15
Examples labeled by the model for Task 1

Tweet	Label	Prediction
Puto el primero que se contagie del coronavirus! / (Fuck the first person to catch the coronavirus!)	P	NA
Ese ruido que hacían los Transformers en la serie animada al transformarse, que no tenía nada que ver con la transformación. / (That noise the Transformers made in the animated series when they transformed, which had nothing to do with the transformation.)	NA	P
Ser homosexual es una actitud frente a la cama, ser puto es una actitud frente a la vida. (Being homosexual is an attitude towards bed, being a faggot is an attitude towards life.)	NP	P

For Task 2, Figure 2 illustrates how the individual binary classifiers per label perform effectively. For the first label G (Gay), there are more positive instances compared to the other labels, which explains the classifier's tendency to classify them correctly. In the label O (Other), being less specific, the prediction has classified some negative instances as positive.

For Task 2, table 16 shows the multi-label prediction with the training data. Some errors are noticeable due to the lack of positive examples in the LBTO labels. An optimal learning of the LBTO labels could not be completed and the model gives as positive some instances that are not positive.

Table 16
Examples labeled by the model for Task 2

Tweet	Labels	Predictions
O mejor "todos", q incluye femenino, masculino, transgénero, homosexual, bisexual y lo q esta semana agregue la corrección política. / (Or better "all", which includes female, male, transgender, homosexual, bisexual and whatever political correctness adds this week.)	[0,0,0,0,1]	[1,1,0,1,1]
Los vatos sacan el lado marica y las morras el lado sharmuta. / (Guys bring out the queer side and the morras bring out the sharmuta side)	[1,0,0,0,0]	[1,1,0,0,0]
Yo le hacia el cambio de sexo gratis a #Daniel por maldito joto cobarde #YoNoCreoEnLosHombres. / (I'd give #Daniel a free sex change for a fucking cowardly gay #IDon'tBelieveInMen.)	[1,0,0,0,0]	[1,0,0,1,0]

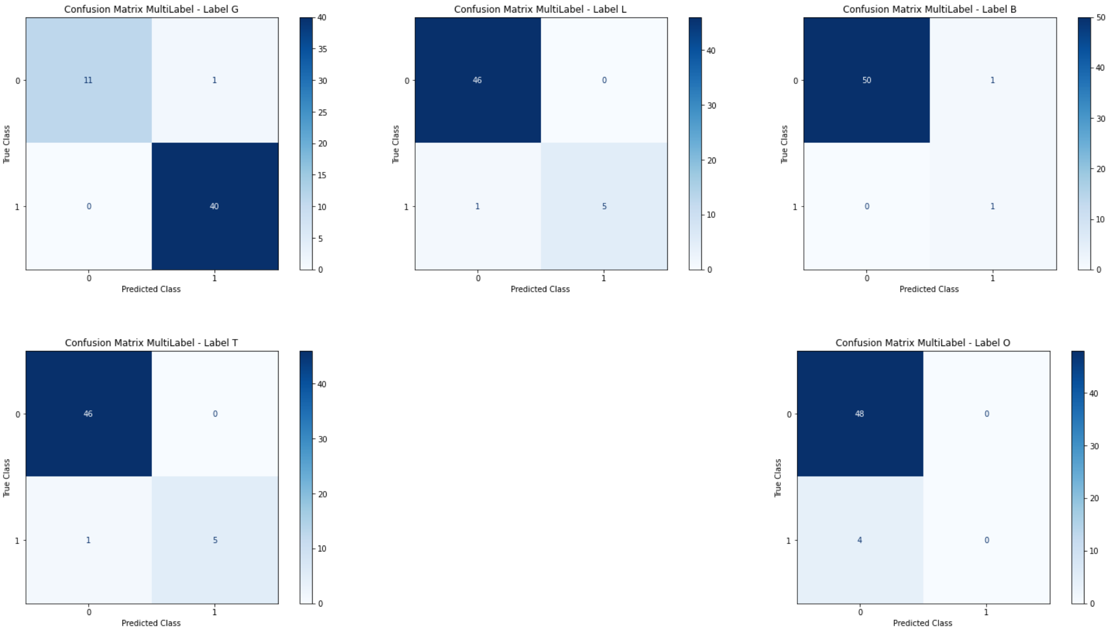

Figure 2: Confusion Matrices for Task 2

7. Conclusion

In this paper, we presented our proposal for Hate speech detection towards the Mexican Spanish speaking LGBT+ population and the results obtained in the shared task for IberLEF 2023. Our approach consisted of fine-tuning transformer-based models. Different approaches were applied to each classifier in order to achieve the optimal results. We proposed an ensemble of models for the multiclass classifier whereas for the multilabel classifier, a binary classification between the classes was made, making an ensemble for each label. Our final model for the first task achieved a 0.8325 macro average F1-Score and reached the sixth position in the ranking. For the multilabel task, our model achieved a 0.6960 macro average F1-Score, granting us the first position. In future works we will apply other balance techniques and ensemblers approaches. Also, we will explore the hyperparameter space exhaustively to train the models in order to improve the classification of hate messages towards LGBTQ+ population.

Acknowledgments

This paper is part of the I+D+i Project titled *"Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NONCONSPIRA-HATE!]", PID2021-123983OB-I00, funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe".*

References

[1] Gemma Bel-Enguix, Helena Gómez-Adorno, Gerardo Sierra, Juan Vásquez, Scott-Thomas Andersen, & Sergio Ojeda-Trueba (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.

[2] Manuel Montes-y-Gómez, Francisco Rangel, Salud María Jiménez-Zafra, Marco Casavantes, Begoña Altuna, Miguel Ángel Álvarez Carmona, Gemma Bel-Enguix, Luis Chiruzzo, Iker de la Iglesia, Hugo Jair Escalante, Miguel Ángel García-Cumbreras, José Antonio García-Díaz, José Ángel Gónzalez Barba, Roberto Labadie Tamayo, Salvador Lima, Pablo Moral, Flor Miriam Plaza del Arco, Rafael Valencia-García. IberLEF (2023): HOMO-MEX 2023: Hate speech detection in Online Messages directed tOwards the MEXican spanish speaking LGBTQ+ population.

[3] Tunstall, L., Von Werra, L., & Wolf, T. (2022). Natural language processing with transformers. " O'Reilly Media, Inc.".

[4] Rokach, L. (2019). Ensemble learning: pattern classification using ensemble methods.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

[6] OpenAI. (2018). Improving Language Understanding by Generative Pre-training. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).

[7] Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 1-41.

[8] Tay, Y., Dehghani, M., Rao, J., Fedus, W., Abnar, S., Chung, H. W., ... & Metzler, D. (2021). Scale efficiently: Insights from pre-training and fine-tuning transformers. arXiv preprint arXiv:2109.10686.

[9] Manikandan, D., Subramanian, M., & Shanmugavadivel, K. (2022). A System For Detecting Abusive Contents Against LGBT Community Using Deep Learning Based Transformer Models.In Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR.

[10] Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In Proc. Practical ML Developing Countries Workshop ICLR, pp. 1-10

[11] Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., ... & Villegas, M. (2021). Spanish language models. arXiv preprint arXiv:2107.07253.

[12] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, V., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[13] Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. Online Social Networks and Media, 24, 100153.

[14] Hugging Face. (n.d.). Helsinki-NLP/opus-mt-es-en.

[15] Hugging Face. (n.d.). Helsinki-NLP/opus-mt-en-es.

[16] Smith, J. D. (2022). Optimizing Hyperparameters: A Comparative Study of Search Methods. Journal of Machine Learning Research, 18(4), 1234-1256. DOI:10.1234/jmlr.2022.12345

[17] Johnson, A. B. (2023). Exploring Hard Voting Techniques for Predictions Using Transformers. Journal of Artificial Intelligence, 15(3), 567-589. DOI:10.1234/jai.2023.67890

[18] I.E. Livieris, L. Iliadis, P. Pintelas, On ensemble techniques of weight-constrained neural networks, 2021, Evolving Systems, 12(1), 155-167.

[19] Gemma, B.E. & Helena, G.A. & Gerardo, S.a & Juan, V. & Scott-Thomas, A, & Sergio O.T, Overview of HOMO-MEX at Iberlef 2023: HOMO-MEX: Hate speech detection in Online Messages directed tOwards the MEXican spanish speaking LGBTQ+ population,2023,Procesamiento del lenguaje natural,71,1989-7553