CIMAT-NLP at HOMO-MEX2023@IBERLEF: Machine Learning Techniques For Fine-grained Speech Detection Task

Erika Rivadeneira-Pérez, María de Jesús García-Santiago and
Cipriano Callejas-Hernández

Mathematics Research Center (CIMAT), Guanajuato, Mexico.

Abstract

With the increasing number of social media users, the number of posts containing Hate Speech (HS) has also increased, leading to various issues. Therefore, it is crucial to develop automatic HS detection systems for social media platforms. In this article, we present some Machine Learning techniques used in HS detection on HOMO-MEX competition task. In particular, our focus is on detecting HS targeted towards the Mexican Spanish-speaking LGBT+ population, addressing the Fine-grained detection problem. This task presents an additional complexity due to its nature as a multi-label problem.

Keywords

HOMO-MEX 2023, Hate Speech

1. Introduction

Over the past years, interest in online Hate Speech (HS) detection and particularly the automatization of this task has continuously grown, along with the social impact of the phenomenon [1]. This has been prompted by the increasing anxieties about the prevalence of Hate Speech on social media, and the psychological and societal harms that offensive messages can cause [2], for instance, in 2016 there was a genocide of Rohingya community in Myanmar as part of an anti-Muslim violence movement made in a Facebook post, in the same study, they reported that posts corresponding to hate speech tends to spread faster than non-hate ones, see [3]. Because of cases like the one in Myanmar, and similar ones, social media platforms have adopted self-imposed definitions, guidelines, and policies for dealing with this particular kind of offensive language. In response, automatic detection of hate speech has become a popular research area in Natural Language Processing (NLP), since what is considered Hate Speech might be influenced by aspects such as the domain of the utterance, its discourse context, and others. In a more deep level study, finding specific targeted groups in Hate Speech discourses is of interest, this is what we referred as fine-grained detection. In this work, we describe our approaches for the HOMO-MEX Hate Speech detection towards the Mexican Spanish speaking LGBT+ population competition track 2: Fine-grained hate speech detection (Multi-labeled) [4].

IberLEF 2023, September 2023, Jaén, Spain

erika.rivadeneira@cimat.mx (E. Rivadeneira-Pérez); maria.garcia@cimat.mx (M. d. J. García-Santiago); cipriano.callejas@cimat.mx (C. Callejas-Hernández)

CEUR Workshop Proceedings (CEUR-WS.org)

1.1. Hate Speech Detection

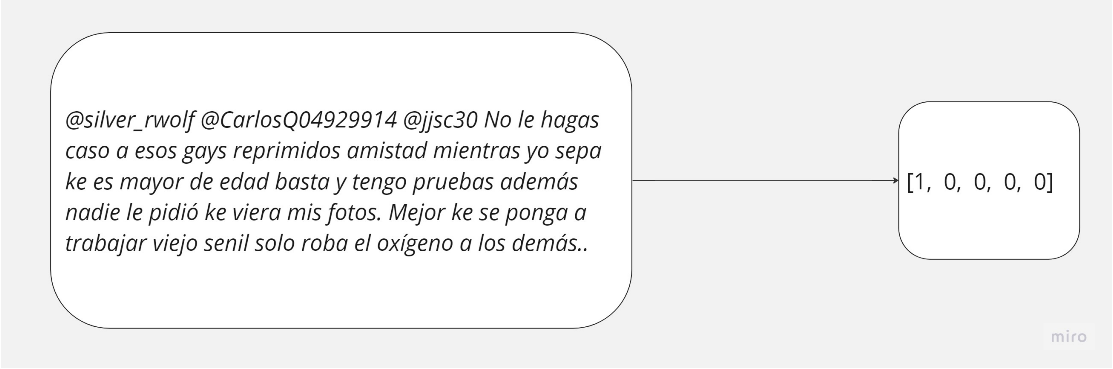
In this work, as in [1] we consider HS as any communication that targets a person or a group based on some characteristics such as race, color, sexual orientation, gender identity, and others. In particular, we are interested in a fine-grained detection, that is, identifying LGBT-specific phobias in each given tweet. What differentiates a hateful speech utterance from a harmless one is probably not attributable to a single class of influencing aspects. While the set of features examined in different works greatly varies the classification methods for this task are mainly focused on supervised learning, which carries an existing bias we shall discuss later. However, we consider a ML approach as a starting point in the approaches described below.

2. Fine-grained hate speech detection track (Multi-labeled).

The various systems developed so far frequently adopt a binary classification framework: given a social media post, a tweet in our case, the system should classify it either as constituting HS or not. In the pioneer work of Davidson et al [5] tweets were primarily filtered as either being offensive language or not, and all offensive tweets subsequently classified as constituting HS or not. Later Qian et al [6] using deep learning techniques distinguish among 40 hate groups, 13 different hate group ideologies (white nationalist, anti-immigration, etc). However as we shall see this fined grained approach depends on there being enough data associated with each sub-type, see [7].

2.1. Corpus Description

The dataset is composed 863 of Mexican Spanish tweets extracted from 2012 to 2022. Each tweet is multi-labeled with a five-entry vector concerning the phobias it contains, either 0 or 1 for each slot, where the first entry is Lesbophobia (L), Gayphobia (G), Biphobia (B), Transphobia (T) and other LGBT+ phobia (O). In figure 1 we have an example of a labeled tweet.



@silver_rwolf @CarlosQ04929914 @jjsc30 No le hagas caso a esos gays reprimidos amistad mientras yo sepa ke es mayor de edad basta y tengo pruebas además nadie le pidió ke viera mis fotos. Mejor ke se ponga a trabajar viejo senil solo roba el oxígeno a los demás..

[1, 0, 0, 0, 0]

Figure 1: Example of tweet and labels of the original dataset.

One challenge associated with this dataset lies in the unbalance distribution of observations across categories. The category pertaining to hate speech towards the Gay population encompasses a substantial majority, accounting for 76% of the observations. In contrast, the tweets categorized as ‘L’ constitute only 7.7%, while ‘B’ category comprise only 1% each, see Figure 2. Consequently, training a classification model with such a dataset leads to an imbalance favoring the Gay population, as there exists a scarcity of observations in the remaining categories (‘L’, ‘B’, ‘T’, and ‘O’). Consequently, the models fail to acquire sufficient knowledge about the characteristics associated with these underrepresented populations.

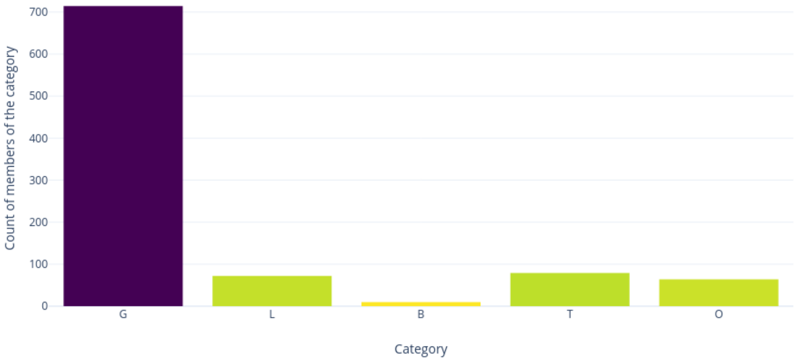
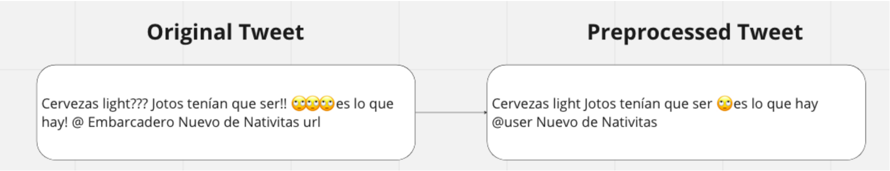


Figure 2: Distribution of tweets into LGBTO categories.

2.2. Corpus Preprocessing and Representation Selection

With the aim of getting homogeneous text data, our preprocessing approach was a follows. See Figure 3 for an example of a preprocessed tweet.



Cervezas light??? Jotos tenían que ser!! 🙄🙄🙄es lo que hay! @ Embarcadero Nuevo de Nativitas url

Cervezas light Jotos tenían que ser 🙄es lo que hay @user Nuevo de Nativitas

Figure 3: Example of a preprocessed tweet.

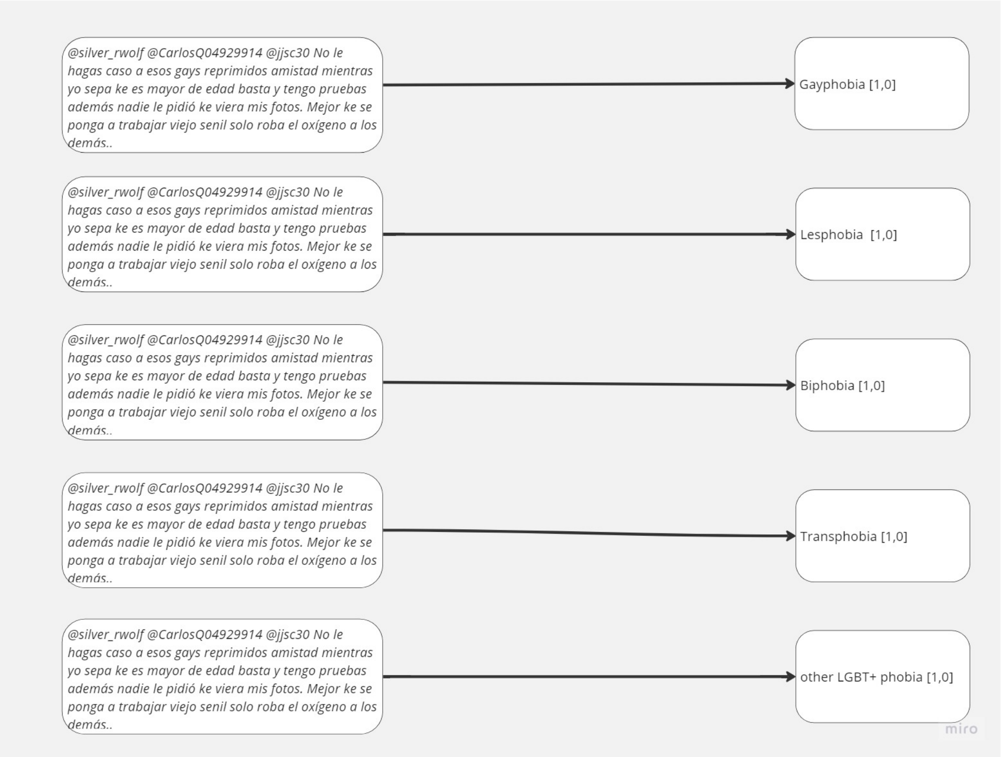
The text transformations applied on the original tweets were the following:

- *Reducing repeated emojis:*
  We consider that the presence of multiple repeated emojis in tweets adds unnecessary noise during the training of our models. Therefore, we reduce the number of repeated emojis to just one occurrence.

- *Removing special characters and URLs:*
  Special characters and URLs in tweets often do not contribute significantly to the classification task and can introduce noise. Thus, we remove them from the text.
- *Mentions substitution:*
  We replace all user mentions with the generic term "@user". This substitution is done to ensure that the model learns the intent of the tweets as a whole, rather than focusing on individual users. By generalizing user mentions, we aim to improve the model's ability to classify LGBT+phobic tweets globally.

2.3. Approach 1.

In this section, we present our first approach for the fine-grained classification problem. This approach is based on the idea of splitting the original problem into several independent binary classification problems, for each category 'L','G','B','T' and 'O' (see Figure 4), and then, using a BOW representation, we classify them with classical machine learning methods.



@silver_rwolf @CarlosQ04929914 @jjsc30 No le hagas caso a esos gays reprimidos amistad mientras yo sepa ke es mayor de edad basta y tengo pruebas además nadie le pidió ke viera mis fotos. Mejor ke se ponga a trabajar viejo senil solo roba el oxígeno a los demás..

@silver_rwolf @CarlosQ04929914 @jjsc30 No le hagas caso a esos gays reprimidos amistad mientras yo sepa ke es mayor de edad basta y tengo pruebas además nadie le pidió ke viera mis fotos. Mejor ke se ponga a trabajar viejo senil solo roba el oxígeno a los demás..

@silver_rwolf @CarlosQ04929914 @jjsc30 No le hagas caso a esos gays reprimidos amistad mientras yo sepa ke es mayor de edad basta y tengo pruebas además nadie le pidió ke viera mis fotos. Mejor ke se ponga a trabajar viejo senil solo roba el oxígeno a los demás..

@silver_rwolf @CarlosQ04929914 @jjsc30 No le hagas caso a esos gays reprimidos amistad mientras yo sepa ke es mayor de edad basta y tengo pruebas además nadie le pidió ke viera mis fotos. Mejor ke se ponga a trabajar viejo senil solo roba el oxígeno a los demás..

@silver_rwolf @CarlosQ04929914 @jjsc30 No le hagas caso a esos gays reprimidos amistad mientras yo sepa ke es mayor de edad basta y tengo pruebas además nadie le pidió ke viera mis fotos. Mejor ke se ponga a trabajar viejo senil solo roba el oxígeno a los demás..

Figure 4: Same tweet as in Figure 1 but treated as five independent binary classification problems.

The procedure of this first approach is as follows:

- **Dataset Splitting:**

Once text data is preprocessed, we divide the training dataset into new subsets, each corresponding to a specific category (L, G, B, T, O), see Figure 4. This division allows us to train classifiers individually for each category. The splitting is performed as follows:

- *Training set*: 90% of the preprocessed data is allocated for training purposes. This large portion of the data ensures that the classifiers have sufficient samples to learn from.
- *Test set*: The remaining 10% of the preprocessed data is used as a test set, providing true labels for evaluation and performance assessment.

For the feature extraction we considered only surface level ones, this is, n-grams for word tokens. This results in a simple TF-IDF weighted BOW representation of the data. Let us remark that this did not change across the five binary classification problems, all parameters were kept equal.

Then, we proceed to train and evaluate classifiers with TF-IDF matrices corresponding to each category of LGBT+phobia. The following classifiers are employed:

- *Random Forest:* This ensemble learning method combines multiple decision trees to improve the classification accuracy [8].
- *Support Vector Machines (SVM):* SVM is a powerful classifier that finds an optimal hyperplane to separate data points into different categories [9].
- *Gaussian processes:* Gaussian processes model the probability distribution over functions and are employed as a probabilistic classifier in our approach [10].

- **Evaluation and Model Selection:**

After training the classifiers, we evaluate the model of each class performance using MAE and F1-score. Based on the evaluation results, we select the best-performing model. See table 2.3.

		Class				
Classifier	**Metric**	**L**	**G**	**B**	**T**	**O**
Random Forest	F1	**0.889957**	**0.872133**	**1.0**	**0.927435**	0.847649
	MAE	**0.08046**	**0.103448**	**0.0**	**0.057471**	0.103448
SVM	F1	0.83109	0.696374	0.982792	0.914643	**0.914643**
	MAE	0.114942	0.21839	0.011494	**0.057471**	**0.057471**
Gaussian Processes	F1	0.83109	0.6701	0.982792	0.914643	**0.914643**
	MAE	0.114942	0.229885	0.011494	**0.057471**	**0.057471**

Table 1
Evaluation of the machine learning methods used in the supervised classification

2.4. Approach 2.

The second approach was again a BOW representation of textual data, but keeping the multilabel, as we still used the main idea in Approach 1, but with a change, in the sense that all classifiers were internally modified instead of considering explicitly five independent binary classification problems.
The preprocess of tweets was kept as in Approach 1, and a variety of traditional ML techniques (Support Vector Machine, RidgeClassifier and Logistic Regression) modified with OnevsRest option were considered with different n-grams. Additionally, a dimensional reduction technique with K best features using the χ^2 function was used.

The OverVSRestClassifier is a strategy that consists in create independent binary classifiers for each label, this means that the classifier fits one specific label versus the second label for classification which is the joined data from the other classes. Consequently, each one of these binary classifiers specializes to classify one phobia. The final output is a straightforward vector of dimension 5 where the one in the entries means the positiveness of the class.

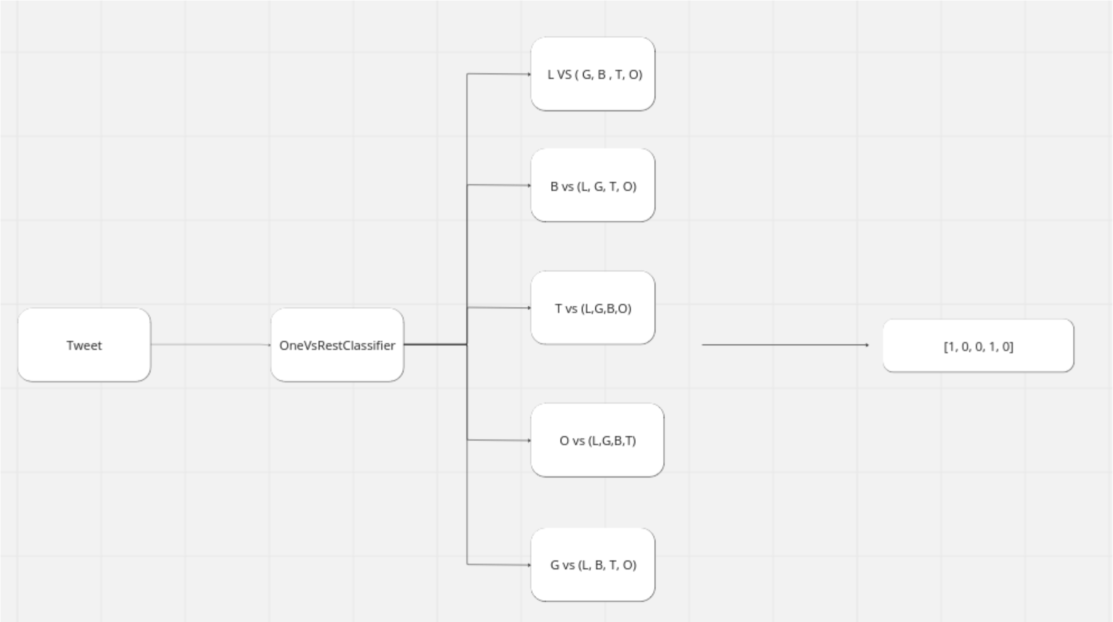


Figure 5: Visualization of the method OneVsRest, in which a multi-label classification problem is turned into five binary ones. For instance, L vs (G,B,T,O) means that tweets labeled as any other but Lesbophobia is are labeled as 0.

2.4.1. Evaluation and Model Selection

We used different models using the given dataset with a partition of 90% train and 10% for test. Then after trial and error, we choose the best model, see 2.4.1 for some tried trials.

Classifier	F1-score	n-grams	Weight	K-best features
SVM	0.641	Unigrams upto 3-ngrams	tf without stopwords	100
SVM	0.595	Unigrams upto 3-ngrams	tf without stopwords	1400
RidgeClassifier	0.593	Unigrams	binary	300
RidgeClassifier	0.593	Unigrams and bigrams	binary	400
RidgeClassifier	0.593	Unigrams up to 3-grams	binary	600

Table 2
Multilabel classifier evaluations

3. Results

The evaluation of the task considered macro-average F1 measure with respect to the positive class for each label with the unweighted mean. We have described our two approaches for *Task 1*, in Table 3 we can see that our first approach reached the third place in this competition.

Approaches	F1-score
I2C-HUELVA	0.6960
Carfer	0.6847
Approach 1	0.6843
Approach 2	0.6579
bayesiano98	0.6812

Table 3
Task 1 results

4. Bias and Ethical Issues

Algorithms are executed automatically and with no human intervention or oversight, opaquely shaping discourse on the internet. It is known that these are useful not only for finding information, but also for providing people with tools to organize and classify knowledge, as well as to take part in social or political discourse. To this point, we emphasize in the intrinsic bias that can be found in this type of datasets, this has consequences, for instance, the content produced by justice organizations is censored or tagged as HS [11], this could lead to the censorship of LBGTQ+ people's attempts to reclaim these words as means for self-expression. On the other hand, the task of classifying millions of offensive tweets is usually crowd sourced, yet it is hard to guarantee quality control using that method. The subjectivity of annotators remains problematic, and it arises from diverging perceptions of what constitutes HS.[7]. Let us remark in this section that two ethical extremes are usually considered during HS detection, either entirely permitting or entirely prohibiting the posting of certain messages, regard as free of speech. Recently new approaches have been proposed, such is the case of quarantining HS [7], situated in between this extremes, where the senders of HS are not censored in a crude unilateral matter, but the recipients of HS are given the option to determine how they wish to handle the HS they have received.

5. Conclusions

Both of our approaches follow the same core, to divide the multi-classification problem into single ones, let us remark here that as seen in Table 3 manually considering independent problems proved to be more efficient that using the internal option of OneVSRest in such classifiers. We hypothesize that this is because we are slightly free to choose either the parameters or classifiers to be used in each binary problem.

In general, our work followed a traditional approach using BOW representation of text data with a Machine Learning classifiers, we believe that the result might be improved if additional features are taken into consideration as proposed in the survey [1]. The ML approach usually yields a good classification performance in binary tasks [12], hence our general approach of splitting. On the other hand, one could suggest the use of pre-trained models such as transformers, but they are known to have limited effectiveness [12]. Finally, in popular opinion, data augmentation can also be considered as well, however in our case we decided not to follow this approach since this can carry the intrinsic bias the train set already presents, see Figure 2.

Acknowledgments

The authors thank Fernando Sánchez-Vega from *Centro de Investigación en Matemáticas* (CIMAT) and *Consejo Nacional de Ciencia y Tecnología* (CONACYT) and its support through the Program "Investigadoras e Investigadores por México" by the project "Desarrollo de Inteligencia Artificial aplicada a la prevención de violencia y salud mental." (ID. 11989, No. 1311) for which the second author is part of.

References

[1] P. Fortuna, M. Domínguez, L. Wanner, Z. Talat, Directions for nlp practices applied to online hate speech detection, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 11794–11805.

[2] E. Heinze, Hate speech and democratic citizenship, Oxford University Press, 2016.

[3] B. Mathew, R. Dutt, P. Goyal, A. Mukherjee, Spread of hate speech in online social media, in: Proceedings of the 10th ACM Conference on Web Science, WebSci '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 173–182. URL: https://doi.org/10.1145/3292522.3326034. doi:10.1145/3292522.3326034.

[4] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S.-T. Andersen, S. Ojeda-Trueba, 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 (2023).

[5] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 512–515.

[6] J. Qian, M. ElSherief, E. Belding, W. Y. Wang, Hierarchical cvae for fine-grained hate speech classification, arXiv preprint arXiv:1809.00088 (2018).

[7] S. Ullmann, M. Tomalin, Quarantining online hate speech: technical and ethical perspectives, Ethics and Information Technology 22 (2020) 69–80.

[8] A. Cutler, D. R. Cutler, J. R. Stevens, Random forests, Ensemble machine learning: Methods and applications (2012) 157–175.

[9] W. S. Noble, What is a support vector machine?, Nature biotechnology 24 (2006) 1565–1567.

[10] C. K. Williams, C. E. Rasmussen, Gaussian processes for machine learning, volume 2, MIT press Cambridge, MA, 2006.

[11] D. O. Thiago, A. D. Marcelo, A. Gomes, Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online, Sexuality & culture 25 (2021) 700–732.

[12] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101. doi:10.18653/v1/W17-1101.