LabTL-INAOE at HOMO-MEX 2024: Distance-based Representations for LGBT+ Phobia Detection

Metztli Ramírez-González, Delia Irazú Hernández-Farías and Manuel Montes-y-Gómez

Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), México

Abstract

In this paper, we describe the LabTL-INAOE participation in the HOMO-MEX 2024 shared task. We propose to use a method based on the distance between a given post and the rest of the instances in the training set to determine whether or not a short comment intends to spread hate speech. For representing texts, we exploited a wide range of schemas ranging from traditional bag-of-words to transformer-based ones. The usefulness of using the distance-based approach was assessed by comparing the results of applying only the text representations for feeding machine learning classifiers. The proposed approach was evaluated in the three subtasks comprised in HOMO-MEX 2024 obtaining competitive results.

Keywords

Hate Speech Detection, LGBT-phobia Detection, Distance-based representations

1. Introduction

Nowadays, approximately 5 million people in Mexico identify with an LGBT+ sexual orientation and gender identity, that is, 1 in every 20 people in the country. Despite this diversity, the latent discrimination and social rejection towards LGBT+ people remain to be present in Mexico. According to official reports, the rate of discrimination of LGBT+ population is twice that of the Non-LGBTI+ population [1]. Any kind of discrimination based on sexual preferences and/or gender identity is defined as *LGBT+phobia* [2]. This is a global problem that has multiple consequences for the LGBT+ community in daily life, such as substance abuse disorders among its members, mental health problems, discrimination in labor markets, denial of access to education and health services, and the lack of human rights [3]. Social networks are a reflection of society, thus there is a growing need to address the detection of LGBT+phobia in them. Timely detection of LGBT+phobic messages can improve content moderation and create safer online environments for users.

This year, in the framework of IberLEF, the *Homo-Mex 2024* shared task was organized [4, 5]. This task is aimed at detecting LGBT+phobia in Mexican Spanish tweets. The detection of LGBT+phobia in Homo-Mex is divided into three tasks:

1. *Hate Speech Detection*: This task aims to predict the label of each tweet. It is a multiclass task in which a tweet can belong to three labels: *a) LGBT+phobic (P)* which includes tweets containing hate speech directed against any person whose sexual orientation and/or gender identity differs from cis-heterosexuality. An example of this class is *"Lo siento, soy muy marica para el dolor"* ("Im sorry, I'm such a fag when it comes to pain"); *b) Non-LGBT+phobic (NP)* comprising tweets mentioning concepts related to the LGBT+ population but without any hate speech intention. An example of this class is *"Estados Unidos levanta la prohibición para que homosexuales donen sangre"*, ("The United States lifts ban on homosexuals donating blood"); and *c) Tweets not related to LGBT+ (NR)* those that are not related in any way to the LGBT+ community. An example of this class is *"Ah v*rga es un duende? Yo pensaba era un alíen asexual"* ("Ah f*ck they're an elf? I thought they were an asexual alien").

IberLEF 2024, September 2024, Valladolid, Spain

*Corresponding author.

metztli.ramirez@inaoep.mx (M. Ramírez-González); dirazuhf@inaoep.mx (D. I. Hernández-Farías); mmontesg@inaoep.mx (M. Montes-y-Gómez)

2. *Fine-grained hate speech detection*: The goal of this multi-label classification task is to predict one or more labels for each individual tweet containing LGBT+ phobic hate speech. Tags are related with various types of hate speech related to LGBT+phobia:
    - *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* for those tweets are those that are not related in any way to the LGBT+ community.
3. *Homophobic lyrics detection*: This is a binary detection task whose objective is to predict whether or not a phrase of a lyrics song contains LGBT+phobic hate speech. It comprises two classes: *a) LGBT+phobic* for lyrics containing hate speech directed against any person whose sexual orientation and/or gender identity differs from cis-heterosexuality, and *b) Not LGBT+phobic* for those lyrics that do not include any hate speech against the LGBT+ population but do mention this community.

This paper describes our participation in the Homo-Mex 2024 shared task. Inspired by the saying "Birds of a feather flock together", we propose a method to detect LGBT+phobic comments that uses a representation based on the distances (with respect to its content) between each post and the rest of the posts from the training set. We also analyze the cases in which this second-order representation causes an improvement in the classification of LGBT+phobia.

This paper is organized as follows. In Section 2, we briefly introduce the solutions made by the participants in *Homo-Mex 2023*, which we consider the most related literature to our proposal. In Section 3, we describe the experimental settings and the obtained results during the developing phase. In Section 4, we present the official results obtained in *Homo-Mex 2024* shared task. Finally, in Section 5, we conclude the paper.

2. Related work

LGBT+phobia on social networks is part of the phenomena covered by *hate speech*, which is defined as a conscious and deliberate public statement intended to denigrate a group of people based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion or political affiliation [6]. Online hate and online extremist narratives have been linked to abhorrent real-world events, including hate crimes and suicides [7]. Detecting hate speech is very challenging since it takes many forms in social media: it can be manifest verbally, non-verbally, and symbolically. Furthermore, hate speech can be expressed in indirect, ambiguous, and metaphorical terms, making its identification even more difficult. It can also be articulated as a negative stereotype that is socially accepted and for which it is not pointed out. Due to the diversity present in hate speech, linguistic analysis is useful but insufficient, because it involves senders, receivers, messages, channels, and interactions, without forgetting its effects and interpretations that feed fear, intimidation, harassment, abuse, and discrimination [8]. For all these reasons, the detection of hate speech is an open problem that must be approached with different solutions.

Diverse shared tasks have been organized with the intention of fostering research on hate speech detection. They have promoted the development of sources of data and as well as motivated the proposal of alternatives to solve different problems in the area of NLP. Among the evaluation campaigns organized to the present day, there is EVALITA focused on the detection of hate speech in Italian [9], There is another task named "Aggression and Gendered Aggression Identification" in three languages

Bangla, Hindi, and English [10], In the framework of SemEval 2019, the task "Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter" focused on Spanish and English [11] was organized. MeOffendES 2021 for the detection of offensive language in Spanish variants [12], and the PAN 2021 focused on identifying hate speech against people based on their race, color, ethnicity, gender, sexual orientation, nationality, religion, etc. [13]. According to [3], there are several efforts to analyze discrimination against the LGBT+ community, such as analyzing data from social networks such as Twitter and Reddit analyzing harassment in cyberspaces, and even generating data on transphobic and homophobic comments.

Homo-Mex is the first shared task focused on detecting LGBT+phobia in Mexican Spanish organized for the first time last year [14]. Mexican Spanish variant is characterized by its particularities in language such as social ingenuity for constructing allegories, insults, and nicknames. It is usually full of ambiguities and contextualization is needed for full understanding. Most of the proposed solutions on the Homo-Mex 2023 involved the use of models based on Transformers and different kinds of data augmentation techniques. Shahiki-Tash et al. [15] used a BERT model and highlighted the importance of performing text preprocessing before using classification models. Rivadeneira-Pérez et al. [16] addressed the multi-label problem with the use of classical methods such as random forests and SVM. Moriña et al. [17] used a transformer ensemble. On the other hand, Marrugo-Tobón et al. [18] and Yigezu et al. [19] exploited data augmentation with different techniques, and used diverse BERT variants for classification. García-Díaz et al. [20] combined embeddings from several Large Language Models (LLMs) in both Spanish and multilingual variations. Rosauro and Cuadros [21] compared classical classification models and Transformers. Macias et al. [22] performed its classification with classic models such as SVM and Bagging Classifier.

3. Experimental Methodology

3.1. Dataset

For training purposes, task organizers provided a dataset for each subtask:

- Task 1: It has a total of 8800 training data, divided into 5482 instances for the *Non-LGBT+phobic* class, 1072 instances for the *LGBT+phobic* class, and 2246 instances for the *irrelevant* class.
- Task 2: It has a total of 1071 training instances, divided into 88 instances marked as *Lesbophobia*, 894 instances marked as *Gayphobia*, 10 instances marked as *Biphobia*, 94 instances marked as *Transphobia*, and 77 instances marked as *Other LGBT+phobia*. It is important to note that in this case, instances can be labeled with more than one label at a time.
- Task 3: It has a total of 984 training instances, divided into 945 *non-LGBT+phobic* instances and 39 or *LGBT+phobic* instances.

3.2. Experiments for the first and second subtasks

Our approach is based on eight stages, which allow us to compare the classification performance obtained with traditional representations (e.g., BoW, contextualized and non-contextualized embeddings) and the distance-based representations obtained from them, which capture the differences (or similarities) in the content of each post with respect to the rest. Figure 1 shows a schematic representation of the phases involved in the proposed approach.

1. **Data preprocessing**: All tweets are lowercase and preprocessed by removing punctuation marks, URLs, line breaks, and stopwords.
2. **First-order representations**:
    - **Traditional representations**: We exploited Bag-of-Words (BoW) using unigrams, bigrams, and trigrams with binary and TF-IDF weighting schemes. Only those terms appearing in at least 10 tweets were considered. Besides, we also filtered out the 500 most representative words for the classes according to the χ^2 statistical measure.

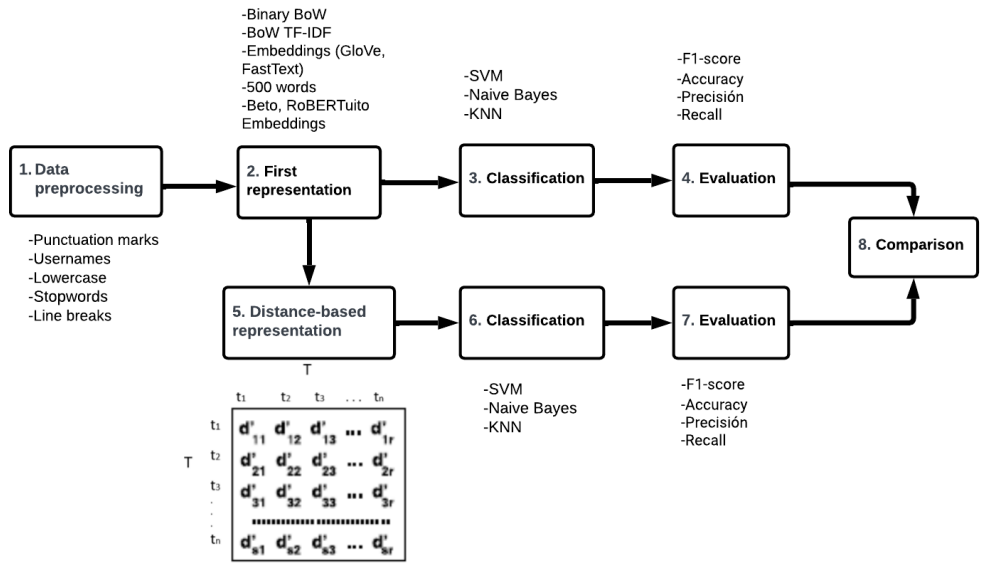

Figure 1: Diagram of the implemented approach.

- **Word embeddings**: We calculated the average vector of each instance considering two pre-trained word embeddings models namely Glove [23] and FastText [24].
- **Transformer embeddings**: We take advantage of the [CLS] vector of two pre-trained models BETO [25] and RoBERTuito [26, 27].

3. **Classification**: During development, a 5-fold cross-validation setting by splitting the training data into two subsets for evaluation purposes using the 80% for training and 20% for validation was used. As classifiers, we use a Support Vector Machine (SVM), Naive Bayes, and k-nearest Neighbors (kNN) for the first subtask. For the second subtask, we used a binary SVM for each label.
4. **Evaluation**: The performance of the classifiers with each representation was evaluated in terms of accuracy, precision, recall, and F1-score.
5. **Second-order representations**: They model each instance, post or song in our case, considering their differences in content with respect to the rest of the elements. Thus, they use these differences as the representation space instead of the conventional characteristics, allowing more general patterns to be found for the distinction between LGBT+ phobic and non-phobic content [28]. To construct the second-order representations we took advantage of the aforementioned first-order representations by comparing each training post with the rest using the *Euclidean distance*, obtaining a square distance matrix. The size of this matrix is determined by the number of training instances (n), being each row the new post representation. In the case of a test instance, its second-order vector representation is obtained by comparing it with all training instances, also obtaining a vector of size n. For **classification** and **evaluation** of the second-order representations, we used the same settings than for the first-order representations.

8. **Comparison**: In this stage, the results of the classifiers based on both representations are analyzed and compared: on the one hand, the first-order representations based on the description of the content of the posts, and, on the other hand, the second-order representations based on the distances (in content) of the posts with respect to the training instances.

3.2.1. Results

Table 1 shows the obtained results by the first-order representations for the *Hate Speech Detection* task. The best result obtained was 0.83 in F1-score terms with the BoW TF-IDF representation with unigrams

and bigrams. For what concerns to the *Fine-grained hate speech detection*, the evaluation was carried out with the macro F1-score. The obtained results are shown in Table 2. In this case, the best performance was achieved by the representation composed of the 500 most relevant words according to χ^2. It is important to note that, a classification rate of 0 was obtained for the class *Biphobia*, which is the one with fewer instances in the dataset.

Table 1
Results of the First-order and Second-order Representations for the First Task.

Representation	Classifier	Accuracy	Precision	Recall	F1	Details
First-order Representations for the First Task						
BoW Binary	SVM	0.8163	0.8392	0.8163	0.8185	Uni-grams
GloVe	SVM	0.6217	0.6976	0.6217	0.6383	Embeddings
FastText	SVM	0.7569	0.8110	0.7569	0.7690	Embeddings
BoW TF-IDF Uni, bi-grams	**SVM**	**0.8323**	**0.8352**	**0.8323**	**0.8333**	Uni, bi-grams
BoW TF-IDF Uni, bi-grams	NB	0.7987	0.7996	0.7987	0.7754	Uni, bi-grams
BoW TF-IDF Uni, bi-grams	kNN	0.6916	0.7188	0.6916	0.6775	Uni, bi-grams
500 words	SVM	0.8274	0.8381	0.8274	0.8291	chi 2
500 words	NB	0.8233	0.8199	0.8233	0.8206	chi 2
500 words	kNN	0.8111	0.8004	0.8111	0.7975	chi 2
BETO Embeddings	SVM	0.7224	0.7772	0.7224	0.7356	Embeddings
RoBERTuito Embeddings	SVM	0.7611	0.8131	0.7611	0.7713	Embeddings
Second-order Representations for the First Task						
BoW TF-IDF Uni, bi-grams	**SVM**	**0.8320**	**0.8359**	**0.8320**	**0.8239**	Uni, bi-grams
BoW TF-IDF Uni, bi-grams	NB	0.6721	0.7154	0.6721	0.5831	Uni, bi-grams
BoW TF-IDF Uni, bi-grams	kNN	0.7976	0.7930	0.7976	0.7946	Uni, bi-grams
500 words	SVM	0.6603	0.8177	0.6603	0.7018	chi 2
500 words	NB	0.5386	0.6331	0.5386	0.5558	chi 2
500 words	kNN	0.7323	0.7694	0.7323	0.7444	chi 2
BETO Embeddings	SVM	0.6841	0.7514	0.6841	0.7007	Embeddings
BETO Embeddings	NB	0.5859	0.6303	0.5859	0.6016	Embeddings
BETO Embeddings	kNN	0.6789	0.6453	0.6789	0.6486	Embeddings
RoBERTuito Embeddings	SVM	0.6250	0.7335	0.6250	0.6489	Embeddings

Regarding the second-order representations, the obtained results for the first task are shown in Table 1. The best performance was achieved when using BoW TF-IDF representation with unigrams and bigrams reaching a 0.82 in F1-score terms. It is interesting to note that, in this case, the experiments performed using a transformer-based representation were (on average) lower than those using traditional schemes. On the other hand, for the second task, the best results obtained were with the BoW TF-IDF representation with unigrams and bigrams 0.49 as shown in Table 2.

Table 2
Results of the First-order and Second-order Representations for the Second Task.

Classifier	Representation	G	L	B	T	O	F1
First-order Representations for the Second Task							
Binary	BoW TF-IDF	0.9350	0.4076	0.0000	0.3716	0.1340	0.3696
Binary	**500 words**	**0.9505**	**0.7304**	**0.0000**	**0.7516**	**0.0917**	**0.5048**
Binary	RoBERTuito Embeddings	0.9116	0.5435	0.0000	0.5985	0.3528	0.4813
Second-order Representations for the Second Task							
Binary	**BoW TF-IDF**	**0.9416**	**0.6643**	**0.0000**	**0.6928**	**0.2006**	**0.4999**
Binary	500 words	0.8683	0.2291	0.1300	0.3514	0.1121	0.3382
Binary	RoBERTuito Embeddings	0.8665	0.4455	0.0500	0.4185	0.3001	0.4161

Once the evaluations of all the experiments were obtained, it was possible to compare both representations. According to the obtained results, we observe no improvement from the first-order representations to the second-order representations in the first task. However, in the second task, there is an improvement from the that there is no improvement from the first-order representations towards the second-order representations at least one of the BoW TF-IDF was used. And when analyzing the data it is possible to observe that the representation based on the use of the 500 most relevant words also has a competitive performance in these tasks. This comparison was crucial to selecting those methods that would be applied for participating in Homo-Mex 2024.

3.3. Experiments for the Third Task

Given that third task consider data of a different domain, we decided to apply a slight variation to our original method. Following we describe the main steps of this new approach.

1. **Data preprocessing**: Elements that indicate punctuation marks, line breaks, and vocal or chorus indicators in the songs were removed and all text was transformed into lowercase.
2. **Data augmentation**: To increase the training data, we added the positive tweets of the second task (as they contain LGBT+phobia content) to the official training data of this one.
3. **First-order representation**: We used a BoW with TF-IDF weights.
4. **Classification**: We used an SVM classifier, and the data was divided into 80% for training and 20% for validation. We used a 5-fold cross-validation setting.
5. **Evaluation**: The F1-score was used as main evaluation metric.
6. **Second-order representation**: From the first-order representation, we calculated the distance matrix for all songs. Each row of this matrix corresponds to the second-order representations of each song.

3.3.1. Results

Table 3 shows the obtained results. The best performance was achieved by the first-order representations with data augmentation with tweets from the second task. As it can be noticed, using data augmentation leads to an improvement in both representations.

Table 3
Results of Data Augmentation with Tweets from the Second Task

Representation	Data	F1
BoW TF-IDF	No data augmentation	0.4882
Distance-based	No data augmentation	0.4880
BoW TF-IDF	**Data augmentation with tweets from the second task**	**0.5439**
Distance-based	Data augmentation with tweets from the second task	0.4927

4. Official Results

According to the results obtained in the previous experiments and their analysis, three representations were selected to be evaluated in Homo-Mex 2024 shared task:

1. **First-order representation:** We selected the BoW with TF-IDF weights as base representation as well as to implement data augmentation for the third task using the tweets from the second task. With this representation, the results obtained in terms of F1-score were: 0.83, 0.89, and 0.49 for the first, second, and third tasks, respectively; these results are shown in Table 4.

2. **Second-order representation:** These representations were constructed from the selected first-order representations. Basically, as explained previously, a distance matrix was constructed, measuring the Euclidean distance between all training posts, with each row being the new representation.
3. **Representation with only the most representative words:** The last representation we evaluated considered only the 500 most representative words according to $\tilde{\chi}^2$.

Table 4 shows the official results obtained in the three subtasks. We also include the ranking position we reached in each case. As it can be noticed, in all cases the participation was based on the use of the Distance-based representation with different configurations for building the matrix.

Table 4
Official results of our participation in the three subtasks of Homo-Mex 2024

Task	Representation	F1	Ranking
Task 1	BoW TF-IDF	0.83	-
	Distance-based	0.85	Fifth
	500 words	0.83	-
Task 2	BoW TF-IDF	0.89	-
	Distance-based	0.88	-
	500 words	0.91	Eighth
Task 3	BoW TF-IDF with data augmentation	0.49	-
	Distance-based	0.48	-
	500 words	0.56	Third

4.1. Discussion and analysis

From the experiments carried out, we observed that the representation based on distances, although obtained similar results to the first-order representations, it presented slight improvements in some cases. In the first task, an improvement from 0.83 to 0.85 was obtained when using the second-order representation. It is also important to stand out that the use of the most representative words demonstrated a competitive performance since it derived the best results in the second and third tasks. In order to analyze the data, word clouds were created for each class and the observations are described below:

- **Non-LGBT+ phobic Class (0):** As seen in Figure 2, the most frequent terms for this class include many terms referring to the LGBT+ community such as "homosexual", "lesbiana", "gay", etc.

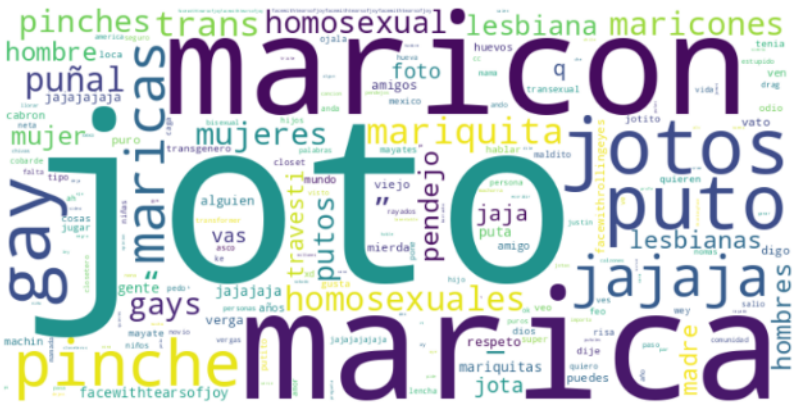

Figure 2: Word cloud for non-LGBT+phobic class.

- **LGBT+ phobia Class (1):** As seen in Figure 3, the most frequent terms for this class are full of derogatory words and many insults referring to the LGBT+ community.

Figure 3: Word cloud for LGBT+phobic class.

- **Irrelevant class (2):** As it is shown in Figure 4, the most frequent terms of this class are varied terms that are not really used within the context of the LGBT+ community, the presence of insults is also seen without them being explicit or referring to the LGBT+ community.

Figure 4: Word cloud for irrelevant class.

When reviewing the most representative terms by class, it is notable that each class has certain particular terms. However, in the tweets of the three classes, there is a presence, to a lesser or greater extent, of the keywords that were used to extract the information, which is why it is not enough to make a distinction between classes.

5. Conclusions

In this paper, we present the LabTL-INAOE participation in the Homo-Mex 2024 shared task. Two approaches were proposed to address the three subtasks of this evaluation campaign. The first one is based on the use of traditional text representations in combination with standard classifiers. The second one attempts to represent each post or song by considering its distance against the rest of the training instances, allowing more general patterns to be found for the distinction between LGBT+ phobic and non-phobic content. A wide range of text representations were used, from classical bag-of-words to transformer-based. These settings were evaluated in both binary and multi-label classification problems, observing an improvement when using the distance-based representations. Although it is a widely used and initial method for representing texts, bag-of-words seems to be very useful and presents

competitive results regarding the use of other more complex methods. The representation based on distance does provide a great improvement in the multiclass problems. Although most experiments involving distance-based representation do not obtain outstanding performance, we have the intuition that by further analyzing and evaluating how to generate distinctive prototypes in this information it will be possible to obtain better results. As future work, we are interested in continue exploring the usefulness of distance-based representations for detecting hate speech content in social media.

References

[1] INEGI, Conociendo a la población LGBTI+ en México, Encuesta Nacional sobre Diversidad Sexual y de Género (ENDISEG) 2021, 2021. URL: https://www.inegi.org.mx/tablerosestadisticos/lgbti/#Poblacion_LGBTI, accedido: 27 de mayo de 2024.

[2] HOMO-MEX, Homo-MEX 24: Hate speech detection towards the Mexican Spanish speaking LGBT+ population, Homo-MEX24. [En línea]. Disponible: https://sites.google.com/view/homomex/home?authuser=0, 2024. [Accedido: 03/05/2024].

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

[4] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, T. Alcántara, M. Soto, C. Macias, Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population, Natural Language Processing 73 (2024).

[5] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[6] S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, H. M. H. López, Internet, social media and online hate speech: Systematic review, Aggression and Violent Behavior 58 (2021) 101608. doi:10.1016/j.avb.2021.101608, art. no. 101608.

[7] N. F. Johnson, R. Leahy, N. J. Restrepo, N. Velásquez, M. Zheng, P. Manrique, S. Wuchty, Hidden resilience and adaptive dynamics of the global online hate ecology, Nature 573 (2019) 261–265. doi:10.1038/s41586-019-1494-7.

[8] M. A. Paz, J. Montero-Díaz, A. Moreno-Delgado, Hate speech: A systematized review, Sage Open 10 (2020) 2158244020973022. doi:10.1177/2158244020973022, art. no. 2158244020973022.

[9] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, Overview of the evalita 2018 hate speech detection task, in: CEUR Workshop Proceedings, volume 2263, CEUR, 2018, pp. 1–9.

[10] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Evaluating aggression identification in social media, in: Proceedings of the second workshop on trolling, aggression and cyberbullying, 2020, pp. 1–5.

[11] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 54–63.

[12] F. M. Plaza-del Arco, M. Casavantes, H. J. Escalante, M. T. Martín-Valdivia, A. Montejo-Ráez, M. Montes, L. Villaseñor-Pineda, Overview of meoffendes at iberlef 2021: Offensive language detection in spanish variants, Procesamiento del Lenguaje Natural 67 (2021) 183–194.

[13] J. Bevendorff, B. Chulvi, G. L. De La Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, E. Zangerle, Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer International Publishing, 2021, pp. 419–431.

[14] 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 toowards the mexican spanish speaking lgbtq+ population, Natural Language Processing 71 (2023).
[15] M. Shahiki-Tash, J. Armenta-Segura, Z. Ahani, O. Kolesnikova, G. Sidorov, A. Gelbukh, Lidoma at HOMOMEX2023@ IBERLEF: Hate speech detection towards the Mexican Spanish-speaking LGBT+ population. The importance of preprocessing before using BERT-based models, in: Proc. Iberian Languages Evaluation Forum (IberLEF 2023), 2023.
[16] E. Rivadeneira-Pérez, M. de Jesús García-Santiago, C. Callejas-Hernández, CIMAT-NLP at HOMOMEX2023@ IBERLEF: Machine Learning Techniques For Fine-grained Speech Detection Task, 2023.
[17] A. J. M. Moriña, J. R. Pásaro, J. M. Vázquez, V. P. Álvarez, I2C-UHU at IberLEF-2023 HOMOMEX task: Ensembling Transformers Models to Identify and Classify Hate Messages Towards the Community LGBTQ, 2023.
[18] D. A. Marrugo-Tobón, J. C. Martinez-Santos, E. Puertas, Natural language content evaluation system for multiclass detection of hate speech in tweets using transformers (2023).
[19] M. G. Yigezu, O. Kolesnikova, G. Sidorov, A. Gelbukh, Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification (2023).
[20] J. A. García-Díaz, S. M. Jiménez-Zafra, R. Valencia-García, UMUTeam at HOMO-MEX 2023: Fine-tuning Large Language Models integration for solving hate-speech detection in Mexican Spanish, 2023.
[21] C. F. Rosauro, M. Cuadros, Hate Speech Detection Against the Mexican Spanish LGBTQ+ Community Using BERT-based Transformers, 2023.
[22] C. Macias, M. Soto, T. Alcántara, H. Calvo, Impact of text preprocessing and feature selection on hate speech detection in online messages towards the LGBTQ+ community in Mexico, in: Proc. of the Iberian Languages Evaluation Forum (IberLEF 2023), 2023.
[23] J. Pennington, R. Socher, C. D. Manning, GloVe: Global Vectors for Word Representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[24] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.
[25] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
[26] J. M. Pérez, J. C. Giudici, F. Luque, pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, 2021. arXiv:2106.09462.
[27] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.
[28] E. Pękalska, R. P. Duin, Dissimilarity representations allow for building good classifiers, Pattern Recognition Letters 23 (2002) 943–956.