Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 361-370 recibido 27-06-2023 revisado 10-07-2023 aceptado 16-07-2023

Overview of HOMO-MEX at Iberlef 2023: Hate speech detection in Online Messages directed Towards the MEXican Spanish speaking LGBTQ+ population

Resumen de HOMO-MEX en Iberlef 2023: Detección de discuros de odio en mensajes online dirigidos hacia la población LGBTQ+ hablante de español mexicano

Gemma Bel-Enguix,[1,4] **Helena Gómez-Adorno,**[2] **Gerardo Sierra,**[1]
Juan Vásquez,[3] **Scott Thomas Andersen,**[3] **Sergio Ojeda-Trueba**[1]
[1]Instituto de Ingeniería (UNAM)
[2]Insituto de Investigaciones en Matemáticas Aplicadas y Sistemas (UNAM)
[3]Posgrado en ciencia e Ingeniería de la computación (UNAM)
[4]Facultat de Filologia i Comunicació (UB)
{gbele, gsierram, sojedat}@iingen.unam.mx, helena.gomez@iimas.unam.mx
{juanmv, stasen}@comunidad.unam.mx

Abstract: The detection of hate speech and stereotypes in online platforms has gained significant attention in the field of Natural Language Processing (NLP). Among various forms of discrimination, LGBTQ+ phobia is prevalent on social media, particularly on platforms like Twitter. The objective of the HOMO-MEX task is to encourage the development of NLP systems that can detect and classify LGBTQ+ phobic content in Spanish tweets, regardless of whether it is expressed aggressively or subtly. The task aims to address the lack of dedicated resources for LGBTQ+ phobia detection in Spanish Twitter and encourages participants to employ multi-label classification approaches.

Keywords: LGBTQ+ phobia, hate speech, machine learning, Twitter.

Resumen: La detección de discursos de odio y estereotipos en plataformas en línea ha suscitado gran atención en el campo del Procesamiento del Lenguaje Natural (PLN). Entre las diversas formas de discriminación, la LGBTQ+fobia es frecuente en las redes sociales, especialmente en plataformas como Twitter. El objetivo de la tarea HOMO-MEX es fomentar el desarrollo de sistemas de PLN que puedan detectar y clasificar contenido LGBTQ+fóbico en tuits en español, independientemente de si se expresa de forma agresiva o sutil. La tarea pretende abordar la falta de recursos dedicados a la detección de la fobia LGBTQ+ en Twitter en español y anima a los participantes a emplear enfoques de clasificación multietiqueta.

Palabras clave: LGBTQ+fobia, discurso de odio, aprendizaje de máquina, Twitter.

1 Introduction

In recent years, significant efforts have been made in the field of Natural Language Processing (NLP) to detect hate speech (Poletto et al., 2021; Lee, Yoon, and Jung, 2018; Kshirsagar et al., 2018; Jarquín-Vásquez, Escalante, and Montes, 2021) and stereotypes (Fraser, Nejadgholi, and Kiritchenko, 2021) in social media platforms.

Although hate speech has been typically related to aggressiveness and the use of slurs, some works (ElSherief et al., 2021) explore the use of implicit hate speech, instead of aggressive expressions. Others works highlight the importance of the context in the interpretation of this content (Dinu et al., 2021).

Traditionally, hate speech has been associated with aggressive behavior and the use

ISSN 1135-5948 DOI 10.26342/2023-71-28 ©2023 Sociedad Española para el Procesamiento del Lenguaje Natural

of derogatory language. However, some studies (ElSherief et al., 2021) have explored the presence of implicit hate speech, which does not rely on explicitly aggressive expressions. Additionally, other research emphasizes the importance of considering the contextual factors when interpreting such content (Dinu et al., 2021).

Among various forms of discrimination and hate speech, LGBTQ+ phobia is prevalent on platforms like Twitter. Several shared tasks have addressed this issue, such as the PAN track at CLEF 2021 (Bevendorff et al., 2021), which utilized a mixed corpus of English and Spanish, and DETOXIS (Taulé et al., 2021), conducted during IberLEF 2021, focusing exclusively on a Spanish corpus.

However, there is a lack of dedicated corpora or shared tasks specifically targeting LGBTQ+ phobia in Spanish Twitter. This is precisely the objective of our corpus, HOMO-MEX, aiming to detect and classify tweets containing LGBTQ+ phobic content, regardless of whether it is expressed aggressively or subtly, using a multi-label classification approach.

The HOMO-MEX task has been introduced at IberLeF 2023 (Jiménez-Zafra, Rangel, and Montes-y-Gómez, 2023; Montes-y-Gómez et al., 2023), a collaborative evaluation campaign for NLP systems focusing on Spanish and other Iberian languages. This initiative is part of the SEPLN congress. The task revolves around the HOMO-MEX corpus, created by the Language Engineering Group.

To compile the corpus, data was collected from Twitter using the API with a geographical filter specific to Mexico. Additionally, a lexicon of LGBTQ+ terms was employed to select the obtained tweets. The corpus underwent a two-step annotation process. In the first step, the tweets were labeled with a three-class classification to detect the presence of LGBT+phobia. Subsequently, the tweets considered positive in the initial stage underwent a multi-label annotation.

This comprehensive corpus and annotation process is the foundation for the HOMO-MEX task, enabling participants to develop NLP systems that detect and classify LGBTQ+ phobic content in Spanish tweets.

The rest of the paper is structured as follows. Section 2 provides a detailed description of the HOMO-MEX 2022 corpus and the evaluation framework utilized in the task. Section 3, presentes various approaches taken by participants to address the problem. Section 4 focuses on the anaysis of the the results obtained by the participating teams. Finally, in Section 5, we summarize the key findings, discuss the implications of the results, and provide potential directions for future research in detecting and combating LGBTQ+ phobia in social media.

2 Homo-Mex 2023 corpus and evaluation framework

For the Homo-Mex task at Iberflef 2023, we created a corpus of tweets in Mexican Spanish that contains nouns indicative of the LGBT+ community. We included slang, slurs and general terminology used to name the members of the LGBT+ collective, without consideration of the polarity of the tweet. First, we created the list of nouns that we collected from social media channels like Twitter, Facebook, Instagram, and TikTok. In addition, we considered any variation of the term that could be present in social media and in the LGBT+ local community. We take into account the extensive gender inflection that characterize the collective, like the use of "efeminization"(efeminización): *puto* could inflect in *pute*, *putx.* A case of efeminization of the corpus is *jota* that derives from *joto*, a noun that does not have gender inflection. Another type of variations contemplated was the appreciative inflection. In Spanish, particularly in the Mexican diaclect, the use of diminutive and augmentative nouns is frequent; in consequence we considered forms like *putito*, *putote*, *putín*, etc. Also, the number inflection was considered.

The final list of lexicon and their approximate translation is in the project's Github.[1]

Having defined the list of search terms, we extracted 706,886 unique tweets provided by Twitter's API. This collection consist of public tweets written in Spanish in Mexico. The dates of this tweets downloaded tweet are 01-01-2012 to 01-10-2022. We annotate 11,000 tweets, half of which are published by unverified accounts and the other half by verified accounts, prior to the monetization of account verification.

We then annotate the tweets to detect LGBT+phobia. In this task, the

[1] https://github.com/juanmvsa/HOMO-MEX

Set	LP	NLP	I	Total
Train	862	4,360	1,778	7,000
Test	477	2,493	1,030	4,000
Total	1,339	6,853	2,808	11,000

Table 1: Size and label distribution for the LGBT+Phobia detection subset.

Set	L	G	B	T	O	Total
Train	72	714	10	79	64	862
Test	34	414	3	38	32	477
Total	106	1,128	13	117	96	X

Table 2: Size and label distribution for the fine-grained classification subset.

tweets are labeled as LGBT+phobic, not LGBT+phobic, and not relevant to the LGBT+ community. The tweets that were labeled as LGBT+phobic were subjected to an additional annotation process, identifying the type of LGBT+phobia. One or more of the following labels could be applied to these tweets, Gayphobia, Lesbophobia, Biphobia, Transphobia, or other types of LGBT+phobia. Although gay is an umbrella term, we asked the annotators to consider Gayphobia as a term encompasing cis-Homosexual men, this was to best distinguish the label from the others. The details of the annotation process are available in (Vásquez et al., 2023).

The label distributions are shown in Tables 1 and 2.

3 Overview of the submitted approaches

Eight teams submitted their working notes detailing their approaches to this shared task. Seven teams participated in Task 1, while Six teams participated in Task 2. The specific teams are detailed in Table 3.

Team - Track 1	Team - Track 2
LIDOMA	Rivadeneira
MarrugoTobon	-
I2C	I2C
mesay	mesay
UMUTeam	UMUTeam
FernandezRosauro	FernandezRosauro
cesar_m	cesar_m

Table 3: Teams that participated in each task.

Tables 4 and 5 show the approaches of each of the teams for tasks 1 and 2 respectively.

Approach	LIDOMA	I2C	mesay	UMUTeam	FernandezRosauro	cesar_m	MarrugoTobon
Transformers	X		X	X	X		X
Single Tran.	X		X		X		
Ensemble Tran.		X		X			X
Traditional ML						X	

Table 4: General approach of each participating team in task 1.

Approach	Rivadeneira	I2C	mesay	UMUTeam	FernandezRosauro	cesar_m
Transformers		X	X	X	X	
Single Transf.						
Ensemble Transf.	X	X	X	X		
Traditional ML	X				X	X

Table 5: General approach of each participating team in task 2.

- *LIDOMA at HOMO-MEX2023@IberLEF: Hate Speech Detection Towards the Mexican Spanish-Speaking LGBT+ Population. The Importance of Preprocessing Before Using BERT-Based Models* (Shahiki-Tash et al., 2023)
  - **Team name: LIDOMA**
  - **Summary:** The team submitted predictions for the first sub-task. As a first step, they converted the labels in the corpus to integers. The authors set a maximum length of 32 tokens for every tweet, and then, they fine-tuned a `bert-base-cased` model with the help of the huggingface library.

- *Machine Learning Techniques For Fine-grained Speech Detection Task*

(Rivadeneira-Pérez, García-Santiago, and Callejas-Hernández, 2023)

- **Team name: Rivadeneira**
- **Summary:** The authors participated in the second sub-task proposed in this shared task. They approached the classification task by dividing the multi-classification problem into single ones. They trained classifiers individually for each category of LGBT+phobia using a training set that consists of 90% of the preprocessed data and a test set that consists of the remaining 10% of the preprocessed data. They used surface level features such as n-grams for word tokens to extract features, resulting in a simple TF-IDF weighted BOW representation of the data. They employed two traditional machine learning classifiers, Random Forest and Support Vector Machines (SVM), to train and evaluate classifiers with TF-IDF matrices corresponding to each category of LGBT+phobia.

- *I2C at IberLEF-2023 HOMO-MEX task: Ensembling Transformers Models to Identify and Classify Hate Messages towards the Community LGBTQ+* (Morano-Morinña et al., 2023)
  - **Team name: I2C**
  - **Summary:** The team submitted results for both sub-tasks. The authors approach the classification task by using an ensemble of classifiers based on transformers. Initially, the authors removed links, usernames, hashtags, and emojis contained in the corpus. Then, with the help of a dictionary, they replaced some terms that are only used in Mexican Spanish for more common terms in other variations of Spanish. Next, they augmented the data in order to balance the corpus. Afterwards, they performed a hyperparameter search for each one of the four pre-trained language models they chose (BETO, RoBERTa, XLM). Finally, for the first sub-task, they obtained their final predictions with a hard-voting technique, while the predictions for the second sub-task were decided by an ensemble approach. One of the most noteworthy results of this paper, is that their confusion matrix shows that the classifiers are not reliable at predicting the class LGBT+Phobia.

- *Natural Language Content Evaluation System For Multiclass Detection of Hate Speech in Tweets Using Transformers* (Marrugo-Tobón, Martínez-Santos, and Puerta, 2023)
  - **Team name: MarrugoTobon**
  - **Summary:** The authors removed the stop words, urls, punctuation marks, and special characters in the corpus. Then, they lowercased the text, followed by an exploratory analysis aimed at extracting keywords, acronyms and abbreviations with the help of n-grams for feature selection. The team implemented a data-balancing phase comprised of two steps: first, they performed a random oversampling. In a second stage, they used cross-validation by random permutation . The classification step was performed with `bert-base-uncased` and `DistilBERT-base-uncased-model`.

- *Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification* (Gemeda-Yigezu et al., 2023)
  - **Team name: Mesay**
  - **Summary:** The Mesay team participated in both sub-tasks. They started with the removal of stopwords, html tags, urls, numbers, and non-alphabetical characters. Then, they lowercased the corpus, and replaced the emojis with a string associated to their meaning. For the first sub-task, they performed an oversampling step to increase the size of the minority class: they randomly duplicated samples

in the minority class. Then, the authors chose to implement BERT-base and RoBERTa-base. The authors find that RoBERTa performed better in the second sub-task, while BERT obtained better results in the first sub-task.

- *UMUTeam at HOMO-MEX 2023: Fine-tuning Large Language Models integration for solving hate-speech detection in Mexican Spanish* (García-Díaz, Jiménez-Zafra, and Valencia-García, 2023)
  - **Team name: UMUTeam**
  - **Summary:** The UMUTeam participated in both sub-tasks. They evaluated different approaches based on the combination of sentence embeddings using ensemble learning and knowledge integration. Specifically, the sentence embeddings were extracted from four Spanish and four multilingual large language models after fine-tuning them for each task separately. The final models were selected using HyperOptSearch with Tree of Parzen Estimators (TPE) and the ASHA Scheduler with the objective of maximizing the macro-weighted F1-score.

- *Hate Speech Detection Against the Mexican Spanish LGBTQ+ Community Using BERT-based Transformers* (C. Fernández-Rosauro and M. Cuadros, 2023)
  - **Team name: FernandezRosauro**
  - **Summary:** The authors utilized both classical machine learning and Transformer-based deep learning models focused on BERT-like architectures to tackle both tracks of the HOMO-MEX task. The baseline models were learned on a TF-IDF matrix generated through the TfidfVectorizer class from the sklearn library. On each task, the vectorizer was fitted on the training dataset and then transformed on both the training and validation datasets. This training and validation TF-IDF matrixes were then used to train and validate the baseline methods. Before the TF-IDF matrices were generated, the tweets followed a processing pipeline, which included a Snowball Stemmer from the NLTK library. The authors used two baseline models: Multinomial Naive Bayes and Linear SVC, both from the sklearn Python library. They used the robertuito-base-uncased model, which achieved the best results in terms of F1-Score (0.84 in Track 1) and macro-average F1-Score (0.68 in Track 2).

- *Impact of Text Preprocessing and Feature Selection on Hate Speech Detection in Online Messages Towards the LGBTQ+ Community in Mexico* (C. Macias et al., 2023)
  - **Team name: cesar_m**
  - **Summary:** The team cesar_m first removed the html entities, line breaks, hashtags, user handles, urls, apostrophes, repeated characters, and alphanumeric words. They finalized the pre-processing step with lowercasing the corpus. Then, they obtained the features with the Bag-of-Words algorithm and the TF-IDF weighting schema. For the first sub-task, the authors trained a linear vector support machine and implemented a bagging classifier from scikit-learn. For the second subtrack, they trained decision trees with a multi-output classifier.

4 Experimental Evaluation and Analysis of the Results

This section reviews the results obtained by the participants of HOMO-MEX at Iberlef 2023. For this purpose, we analyze and compare the submitted solutions' performance on the test set. We used the F1-macro-score as the primary performance measure to rank all the participants. We launched a Codalab competition[2] to manage the shared

[2] https://codalab.lisn.upsaclay.fr/competitions/10019

task stages and compute the performance metric for all submissions.

As baseline approach, for Task 1, we use `bert-base-multilingual-cased` without any pre-processing. For Task 2, we use the `twitter-xlm-roberta-base-sentiment` model from Hugging Face (Barbieri, Espinosa Anke, and Camacho-Collados, 2021).

Table 6 summarizes the results obtained by each team and our baseline in task 1 of HOMO-MEX shared task. We report the F1 score in each class, the macro F1 score, and the accuracy. In this edition of the HOMO-MEX shared task, the approach submitted by the (C. Fernández-Rosauro and M. Cuadros, 2023) team outperformed all the other approaches and the baseline. The (C. Fernández-Rosauro and M. Cuadros, 2023) team submitted an approach based on the RoBERTuito model (Pérez et al., 2021). RoBERTuito is specifically designed for social media text in Spanish and trained on a large dataset of 500 million Spanish tweets. This language specificity allows RoBERTuito to better capture the nuances and characteristics of Spanish text, including the context and language patterns related to LGBTphobia. The team used the pysentimiento library to preprocess tweets before feeding them into the RoBERTuito model. This additional preprocessing step could have helped to clean and prepare the text data specifically for the characteristics of RoBERTuito, potentially improving the model's ability to understand and classify LGBTphobic content accurately. Also, the team addressed the class imbalance issue, which specifically affects the identification of LGBT+phobic (P) tweets, by adding a class weight dictionary during the training phase. By assigning different weights to each class, with a higher weight for the minority class (LGBT+phobic tweets), the model can give more attention and importance to the minority class, potentially improving its performance in identifying these instances correctly.

The second-best approach proposed by the (García-Díaz, Jiménez-Zafra, and Valencia-García, 2023) team explored different sentence embedding combination strategies using knowledge integration and ensemble learning. The team tested various Spanish and multilingual LLMs, including BETO, MarIA, AlBETO, DistilBETO, BERT, MdeBERTA, TwHIN, and XLM. They evaluated different approaches based on the sentence embeddings extracted from these LLMs after fine-tuning them separately for each task. The knowledge integration strategy implemented involved training a multi-input neural network with all sentence embeddings introduced simultaneously. The specific architectural choices and hyperparameter settings likely contributed to its high performance in the LGBT+phobia detection.

We employ two metrics, Maximum Possible Accuracy (MPA) and Coincident Failure Diversity (CFD), to assess the complementary and diversity of the predictions provided by the different approaches (Tang, Suganthan, and Yao, 2006). Table MPA measures the accuracy of the classifications by calculating the ratio of correctly classified instances to the total number of instances. For an instance to be considered correctly classified, at least one team must assign the correct label to it. By using the MPA metric, we can identify instances that have been misclassified by all teams.

In contrast, the CFD metric ranges from a minimum value of 0 to a maximum value of 1 (Kuncheva and Whitaker, 2003). A CFD value of 0 indicates that either all classifiers are always correct or all classifiers are consistently incorrect. Conversely, a CFD value of 1 suggests that, at most, one classifier will fail for any randomly chosen instance. Equation 1 defines the CFD metric.

$$CFD = \begin{cases} 0, & p_0 = 1.0; \\ \frac{1}{1-p_0} \sum_{i=1}^{L} \frac{L-i}{L-1} p_i & p_0 < 1 \end{cases} \tag{1}$$

The results of the MPA and CFD metrics are presented in Table 7, where the proposed approaches are grouped based on their shared characteristics. We have created five groups: "All Teams," which consists of all participating teams; "Transformers" approaches, comprising teams such as LIDOMA, I2C, mesay, UMUteam, FernandezRosauro, and Marrugo; We subdivided the teams in transformers approaches in single transformers and ensemble transformers to differentiate when teams are using an ensemble of transformers models; "Traditional ML" based approach which consists of only one team, cesar_m.

Team	F1-score (P)	F1-score (NP)	F1-score (NA)	Macro F1-score
FernandezRosauro	**0.7122**	**0.9153**	**0.9020**	**0.8432**
UMUTeam	0.7046	0.9182	0.9036	0.8421
I2C	0.6868	0.9148	0.8960	0.8325
mesay	0.6338	0.8909	0.8653	0.7967
Marrugo	0.5898	0.8639	0.8578	0.7705
cesar_m	0.5829	0.8874	0.8203	0.7635
LIDOMA	0.5679	0.8675	0.7625	0.7326
baseline	0.7169	0.9155	0.8845	0.8390

Table 6: Result summary for the HOMO-MEX shared task on Task 1.

Approach	Best accuracy	MPA	CFD	Number of systems
All teams	0.8880	0.9623	0.1355	7
Transformers	0.8858	0.9547	0.1299	6
Single Transformer	0.8858	0.9455	0.1555	4
Ensemble Transformer	0.8880	0.9210	0.0765	2
Traditional ML	0.8403	0.8403	-	1

Table 7: MPA and CFD comparison results among the different proposed approaches for task 1.

All teams, including Transformer-based and Traditional ML approaches, achieved relatively high MPA scores. The overall MPA for all teams is 0.9623. The Transformer-based approaches, including Single Transformer and Ensemble Transformer, achieved MPA scores of 0.9547 and 0.888, respectively. The Traditional ML approach obtained a slightly lower MPA score of 0.8403. Based on the MPA scores, the Transformer-based approaches performed slightly better than the Traditional ML approach. On the other hand, the CFD scores suggest that the Single Transformer approach exhibited higher diversity in predictions compared to the other approaches. Further analysis is needed to understand the reasons behind the differences in performance and explore each approach's potential strengths and weaknesses. Some weaknesses in the detection may rely on the linguistics ambiguity and polysemous nouns used to name the LGBT+ community. For example, in the tweet *Mis datos están bien maricas* that could be translated to *My data is faggot*; this use of the word *marica* is notably uncommon. Also, in *Así me vería si fuera vestida* (*I would look like that if I was dressed*) is ambiguous to determine if the author refers to the way he looks in a dress or if he would look like that if he were a cross-dresser. Any of the task participants did not assert these examples cited.

Finally, Table 8 reports the results obtained by the teams in task 2.

5 Conclusions

This paper presents the design and findings of the HOMO-MEX shared task, which was held in conjunction with IberLef 2022. HOMO-MEX focuses on detecting LGBT-phobia, including a more detailed subtask of distinguishing between gay, lesbian, bisexual, and transgender phobia. This inaugural edition of the task yielded interesting results.

The first task, LGBT-phobia detection, involved a three-class classification and achieved moderately satisfactory outcomes. However, the second task, which entailed multi-label identification, received lower performance scores.

In general, most teams employed various Transformers models as the foundation for their methodologies. Only one team preferred to use classical ML models. This shows how transformers have has become the new standard for dealing with many linguistic problems.

Future endeavors will encompass several challenging aspects, such as expanding the corpus to include other Latin American Spanish dialects. Additionally, further investigation is required to explore the potential biases introduced by annotator sociodemographic factors. Another issue to investigate is the impact of the author in the statements, this is, whether the addresser belongs or not

Team	Macro F1-score
Rivadeneria	**0.6843**
I2C	0.6960
mesay	0.6868
UMUteam	0.6338
cesar_m	0.5679

Table 8: Result summary for the HOMO-MEX shared task on Task 2.

to the LGBTQ+ community can change the polarity and intention of the message.

Acknowledgements

This paper has been supported by PAPIIT projects IT100822, TA101722, and CONAHCYT CF-2023-G-64. Also, we thank Alejandro Ojeda Trueba for the creation of the HOMO-MEX presentation image. GBE is supported by a grant from the Ministry of Universities of the Government of Spain, financed by the European Union, NextGeneration EU (María Zambrano program).

References

Barbieri, F., L. Espinosa Anke, and J. Camacho-Collados. 2021. XLM-T: A multilingual language model toolkit for twitter. *CoRR*, abs/2104.12250.

Bevendorff, J., B. Chulvi, G. Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al. 2021. Overview of pan 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 419–431. Springer.

C. Fernández-Rosauro and M. Cuadros. 2023. Hate speech detection against the mexican spanish lgbtq+ community using bert-based transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).*

C. Macias, M. Soto, T. Alcántara, and H. Calvo. 2023. Impact of text preprocessing and feature selection on hate speech detection in online messages towards the lgbtq+ community in mexico. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).*

Dinu, L. P., I.-B. Iordache, A. S. Uban, and M. Zampieri. 2021. A Computational Exploration of Pejorative Language in Social Media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.

ElSherief, M., C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. D. Choudhury, and D. Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *CoRR*, abs/2109.05322.

Fraser, K. C., I. Nejadgholi, and S. Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model.

García-Díaz, J., S. Jiménez-Zafra, and R. Valencia-García. 2023. Umuteam at homo-mex 2023: Fine-tuning large language models integration for solving hate-speech detection in mexican spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).*

Gemeda-Yigezu, M., O. Kolensikova, G. Sidorov, and A. Gelbukh. 2023. Transformer-based hate speech detection for multi-class and multi-label classification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).*

Jarquín-Vásquez, H., H. J. Escalante, and M. Montes. 2021. Self-contextualized attention for abusive language identification. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 103–112, Online, June. Association for Computational Linguistics.

Jiménez-Zafra, S., F. Rangel, and M. Montes-y-Gómez. 2023. Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF*

2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org.

Kshirsagar, R., T. Cukuvac, K. R. McKeown, and S. McGregor. 2018. Predictive embeddings for hate speech detection on twitter. *CoRR*, abs/1809.10644.

Kuncheva, L. and C. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 05.

Lee, Y., S. Yoon, and K. Jung. 2018. Comparative Studies of Detecting Abusive Language on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106, Brussels, Belgium. Association for Computational Linguistics.

Marrugo-Tobón, D., J. Martínez-Santos, and E. Puerta. 2023. Natural language content evaluation system for multiclass detection of hate speech in tweets using transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).*

Montes-y-Gómez, M., F. Rangel, S. Jimeńez-Zafra, M. Casavantes, B. Altuna, M. Álvarez Carmona, G. Bel-Enguix, L. Chiruzzo, I. de la Iglesia, H. Escalante, M. Garciá-Cumbreras, J. Garciá-Diáz, J. Gonzalez Barba, R. Labadie Tamayo, S. Lima, P. Moral, F. Plaza del Arco, and R. Valencia-Garciá, editors. 2023. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org.*

Morano-Morinña, A., J. Román-Pásaro, J. Mata-Vázquez, and V. Pach'øn-Álvarez. 2023. I2c at iberlef-2023 homomex task: Ensembling transformers models to identify and classify hate messages towards the community lgbtq+. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).*

Pérez, J. M., D. A. Furman, L. A. Alemany, and F. Luque. 2021. Robertuito: a pretrained language model for social media text in spanish. *CoRR*, abs/2111.09453.

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Rivadeneira-Pérez, E., M. García-Santiago, and C. Callejas-Hernández. 2023. Machine learning techniques for fine-grained speech detection task. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).*

Shahiki-Tash, M., J. Armenta-Segura, Z. Ahani, O. Kolesnikova, G. Sidorov, and A. Gelbukh. 2023. Lidoma at homomex2023@iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).*

Tang, E., P. Suganthan, and X. Yao. 2006. An analysis of diversity measures. *Machine learning*, 65(1):247–271.

Taulé, M., A. Ariza, M. Nofre, E. Amigó, and P. Rosso. 2021. Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish. *Procesamiento del Lenguaje Natural*, 67:209–221.

Vásquez, J., S. T. Andersen, G. Bel-Enguix, H. Gómez-Adorno, and S.-L. Ojeda-Trueba. 2023. Homo-mex: A mexican spanish annotated corpus for lgbt+phobia detection on twitter. In *Proceedings of The 7th Workshop on Online Abuse and Harms (WOAH)*, Toronto, Canada, July. Association for Computational Linguistics.

A Annex 1: Data Collection Ethics

We collect tweets from the social media platform Twitter using the Twitter API. This API permits the collection of tweets that have been publicly posted. The authors of the tweets are not notified of their tweets participation in this study, however this process is in accordance to Twitter's privacy policy. We ensure adherence to the requirements Twitter sets for use of this API.

The tweets collected are based on tagged metadata. All scraped tweets had geolocation tags in Mexico, and a language tag for

Spanish. These tweets are supposedly provided to us randomly by the API, we assume a variety of author demographics are represented, such as variations in race, nationality, and socioeconomic background. However these are not facts that we can verify.

We selected annotators that self identified as members and non-members of the LGBT+ community. They were informed of the purpose of the study and the harmful nature of some of the tweets they would be labeling, and were informed that they could stop participation in the study at time if they did not wish to continue.