DSVS at HOMO-MEX24: Multi-Class and Multi-Label Hate Speech Detection using Transformer-Based Models

Sergio Damián[1,*], David Vázquez[1], Edgardo Felipe-Riverón[1] and Cornelio Yáñez-Márquez[1]

[1]Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), México

Abstract

The present work describes the participation of the DSVS team in the HOMO-MEX shared task at IberLEF 2024 on detecting hate speech in online messages and music lyrics targeting the LGBTQ+ community, written in Mexican Spanish. The study addressed all three proposed tracks: Track 1 involves identifying LGBTQ+ categories (multiclass); Track 2 focuses on fine-grained hate speech detection (multi-labeled); and Track 3 involves homophobic lyrics detection (binary task). Through an exploration of the datasets, we employ various BERT-based models. Our team's best submission secured the 4th position for Track 1, the 3rd position for Track 2, and the 9th position for Track 3.

Keywords

Hate Speech Detection, LGBTQ+ phobia, Transformers, Large Language Models, Natural Language Processing

1. Introduction

LGBT+phobia refers to a wide range of negative attitudes, prejudices, and discriminatory behaviors directed towards individuals with non-normative gender identities and sexual orientations.

This behavior has deep cultural, social, and religious foundations and can be present in individuals, institutions, and countries. For instance, data from ILGA World (International Lesbian, Gay, Bisexual, Trans, and Intersex Association) [1] reveals that 60 UN member nations criminalize consensual same-sex acts by law, with two more doing so de facto which are denoted in Figure 1.

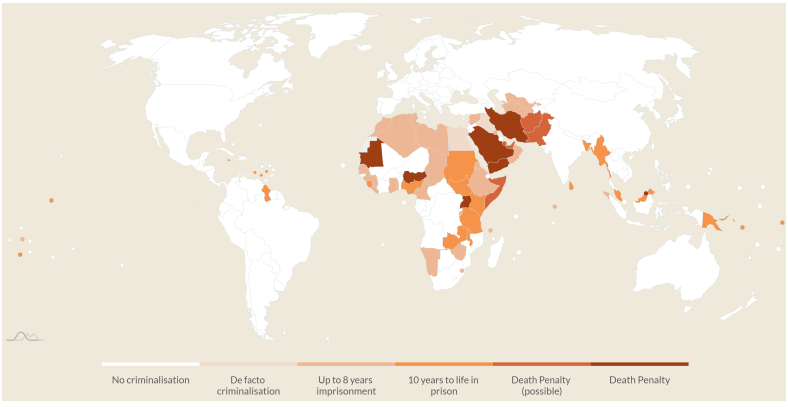

Figure 1: Criminalization of consensual same-sex sexual acts, UN member states according to [1].

LGBT+phobic behaviors have significant repercussions in society. In Italy, several factors converge to create a highly unfavorable environment for LGBT+ people, including cultural and religious influences. People belonging to sexual minorities can feel rejected by their religious community due to their sexual orientation. Furthermore, in highly religious families, parents often tend to abandon their LGBT+ children [2, 3, 4].

IberLEF 2024, September 2024, Valladolid, Spain

sdamians2019@cic.ipn.mx (S. Damián); dvazquezs2019@cic.ipn.mx (D. Vázquez); edgardo@cic.ipn.mx (E. Felipe-Riverón); cyanez@cic.ipn.mx (C. Yáñez-Márquez)

https://github.com/sdamians (S. Damián); https://github.com/Hiram02 (D. Vázquez)

The exclusion of these individuals from society can lead to family breakdown. This rejection can also weaken social bonds and contribute to greater social disintegration. Additionally, LGBT+ individuals who belong to groups typically associated with a low tolerance for nonheteronormative practices, such as certain religious groups, may experience even greater segregation. For example, members of the Italian Catholic LGBT+ organization risk being isolated from both their religious community and the LGBT+ community [5].

The ramifications of marginalizing minority groups extend beyond social dynamics, significantly impacting a nation's economic landscape. In South Africa, discriminatory employment practices and limited job opportunities for diverse sexual orientations and gender identities have led to substantial economic losses [6]. Every year, South Africa bears the burden of approximately $316.8 million due to this problem alone. Moreover, health disparities cost the country between $3.2 billion and $19.5 billion annually, driven by elevated rates of illness and mental health conditions necessitating expensive treatments. Additionally, the prevalence of sexual violence and other forms of victimization among LGBT+ individuals presents another formidable economic challenge. The annual economic cost associated with sexual violence against LGBT+ adults is estimated at $64.8 million.

The situation in Mexico is equally alarming, with authorities often failing to protect the LGBT+ community from violence and discrimination. Several reports indicate that police and military personnel are sometimes responsible for harassing and assaulting transgender women. Numerous acts of violence, harassment, and even false arrests have been perpetrated by Mexican authorities against transgender women [7, 8, 9, 10, 11]. This situation is further aggravated by family rejection, which has led to up to 70% young transgender Mexicans having fled or been thrown out of their homes [9].

These examples highlight how attacks on the LGBT+ community not only impact the community itself but also have a profound influence on society at large. Hence, it is crucial to address the issues linked to discrimination and hatred faced by the LGBT+ community.

Despite significant advances against discrimination against the LGBT+ community, the issue of hate speech persists. This reality underscores the necessity for initiatives such as the shared HOMO-MEX 24 task: Hate speech detection towards the Mexican Spanish-speaking LGBT+ population [12], which was organized as part of IberLEF 2024 [13]. These efforts aim to address and mitigate the harmful impact of hate speech directed at the LGBT+ community, emphasizing the importance of combating discrimination and fostering inclusivity.

Our team participated in all three tracks, conducting experiments with various datasets and Large Language Models (LLMs) for classification, which were fine-tuned for each track. These LLMs include Spanish models such as BETO [14], MarIA [15], Alberti [16], Bertin [17], RoBERTuito [18], and TwHIN-BERT [19]. The source code for this work can be found at https://github.com/sdamians/homomex24.

2. Related Work

The HOMO-MEX shared task is being held at IberLef for the second year. In previous years, the tasks involved analyzing and experimenting with texts containing the context of the LGBT+ community in Mexico [20]. Various approaches were developed during the last event, employing different techniques utilizing a dataset created for this purpose [21]. The best approach showcased the utilization of fine-tuned transformer-based models, including BETO, MDeberta [22], and RoBERTuito, with the latter being identified as the best-performing model, yielding the most favorable results [23]. Some additional approaches that were submitted during the last event included ensemble models utilizing sentence embeddings from transformer-based models [24], the single use of the BERT [25] model, the feature extraction through text translation from Spanish to English [26], the utilization of traditional machine learning models such as Decision Trees and Support Vector Machines with TF-IDF representations [27], among others.

The approach implemented in this work was inspired by the strategy used by the winning team in last year's event. Multiple transformer-based models were fine-tuned to get the best results.

RoBERTuito was the most suitable model for tracks 1 and 2 this year because it was pre-trained on Twitter-like texts and other forms of social media language. For analyzing song lyrics, Alberti was chosen, a model which had been trained in poetry texts and other metaphorical data. This choice could extract contextual information about certain words that could convey indirect insults to the LGBT+ community.

3. Dataset description

3.1. Track 1: Hate Speech Detection (Multi-Class Classification)

The training dataset for this track includes inputs written in Mexican Spanish from social media. There are three distinct labels in this dataset: LGBT+phobic (P), not LGBT+phobic (NP), and not LGBT+related (NR). During the cleaning phase, duplicate inputs were eliminated, prioritizing the preservation of the majority class presence. In cases of a tie, the class associated with the most recent occurrence was selected. Furthermore, user mentions were kept because their absence resulted in the generation of duplicate inputs with different class labels. This process culminated in a dataset containing a total of 17,943 inputs. For this dataset, the following procedures were also implemented:

- Hashtag words were split according to the camel case syntax.
- User mentions were replaced with the token "usuario". If more than one user mention was present subsequently, the token "usuarios" was used instead.
- Tokens representing URLs were removed.
- Line breaks were transformed into spaces.
- Space reduction was applied.
- Emoji tokens were converted into word representations.
- HTML entities were decoded.
- Non-Spanish characters were removed (e.g. Asian characters).
- Special Twitter tokens such as "MD" and "RT" were eliminated.

In the process of cleaning and analyzing the data, the analysis revealed a significant class imbalance in the dataset, with the "P" label representing the minority class (LGBT+ phobia content related). To tackle this issue, three different versions of the cleaned dataset were proposed. The first version followed the data preprocessing described previously, the second version removed emojis, and the third version used the predefined Robertuito cleaning method, which involved replacing emojis and tweet-specific elements with predefined tokens. Table 1 illustrates the distribution of inputs across the "P", "NP" and "NR" classes in both the training and validation datasets. For validation and test datasets, duplicated inputs were not identified.

Table 1
Dataset Description for Track 1. The "Original Train" column represents the total number of inputs received in the train dataset, encompassing all instances prior to the removal of repeated inputs.

Label	Numerical Label	Original Train	Train	Val	Test	Total
LGBT+phobic (P)	0	1072	1067	862	?	1934
Not LGBT+phobic (NP)	1	5482	5455	4360	?	9826
Not LGBT+related (NR)	2	2246	2221	1778	?	4015
Total	-	8800	8743	7000	2200	17943

3.2. Track 2: Fine-grained Hate Speech Detection (Multi-Labeled Classification)

In this dataset, detailed categorization is achieved by using fine-grained labels. These labels consist of Lesbophobia (L), Gayphobia (G), Biphobia (B), Transphobia (T), Other LGBT+phobia (O), and Not

LGBT+related (NR). The entries may be associated with one or multiple fine-grained labels, illustrating the diverse nature of hate speech directed towards the LGBT+ community. It was observed that no entries were classified under the NR label, but it was still accounted for during the training process as required in the track description. As indicated by Table 2, most of the data was categorized under the Gayphobia (G) label, with minimal representation of combinations of other labels. Emojis were considered during the data preprocessing phase to extract relevant information from the inputs and non-Spanish characters were removed. Considering these factors, three dataset versions were created, following the approach used in the preceding track. These versions included: one with emojis, one without emojis, and a third implementing the Robertuito method, containing 2199 entries.

Table 2
Dataset Description for Track 2.

	L	G	B	T	O	NR	Original Train	Train	Val	Test	Total
		✓					830	828	658	?	1486
				✓			66	66	58	?	124
	✓						57	57	46	?	103
					✓		33	33	30	?	63
		✓			✓		28	28	24	?	52
	✓	✓					20	20	16	?	36
				✓	✓		10	10	4	?	14
		✓		✓			7	7	7	?	14
	✓			✓			5	5	4	?	9
			✓				4	4	4	?	8
		✓	✓				3	3	3	?	6
	✓	✓		✓			2	2	2	?	4
	✓	✓		✓	✓		1	1	1	?	1
	✓	✓			✓		1	1	1	?	1
		✓		✓	✓		1	1	1	?	1
	✓		✓	✓	✓		1	1	1	?	1
			✓		✓		1	1	1	?	1
	✓	✓	✓	✓	✓		1	1	1	?	1
Total							1071	1069	862	268	2199

3.3. Track 3: Homophobic Lyrics Detection (Binary Classification)

As shown in Table 3, the LGBT+phobic (P) class has significantly fewer samples compared to the opposing class, Not LGBT+phobic (NP). The dataset had inputs that exceeded the maximum token limit allowed by the models used in this work. The entries with more than 128 tokens were segmented to increase the training data volume and balance the class proportions. Additionally, non-Spanish text segments of lyrics were translated using the Google Translator API. Song annotations enclosed within brackets and comments delineated by parentheses were removed. Artist names were tagged with *user* to enhance data quality, and repeated words and duplicate inputs were also eliminated. This resulted in a dataset containing a total of 1,896 inputs.

Table 3
Dataset Description for Track 3. Data augmentation by splitting long entries was considered for the training step. Duplicate inputs were removed.

Label	Numerical Label	Original Train	Train	Val	Test	Total
LGBT+phobic (P)	0	39	146	40	?	186
Not LGBT+phobic (NP)	1	945	904	560	?	1464
Total	-	984	1050	600	246	1896

4. Methodology

BETO, TwHINBERT, BERTIN, and RoBERTuito models were fine-tuned for tracks 1 and 2. For track 3, BETO, MarIA, and Alberti were fine-tuned. During experimentation, we varied diverse parameter values, such as the number of epochs, batch size, dropout rate, learning rate, and weight decay, shown in Table 4. In addition, a polynomial learning rate scheduler was used. Class weights were integrated into the loss functions to address the class imbalance in each dataset. Despite attempting techniques like lowercasing the texts and freezing layers of the language models, there were no significant improvements. However, some models showed improved performance when emoji tokens were removed, while others performed better with the presence of emoji tokens. Table 5 presents the results obtained for the evaluation dataset per each track. Robertuito got the best results for track 1, BETO for track 2, and MarIA for track 3. For the first two tracks, emoji tokens were significant for the solution, although BETO without emoji got the best results for Hamming Loss and the Exact Match Ratio (MR). For the third track, MarIA outperformed the other two implemented approaches significantly.

Table 4
Parameter variation for each experiment developed. Generally, low dropout rates and a value of zero for weight decay obtained the best results.

Parameter	Range / Values
Epochs	10,15,20,25
Batch size	8, 16
Dropout rate	0.0, 0.05, 0.1, 0.2
Learning rate	2e-5, 4e-5, 5e-5
Weight decay	0.0, 0.001, 0.01
AMSGrad	True,False
Epsilon	1e-8,1e-10

Table 5
Evaluation results for the best experiments based on the combination of LLM and preprocessed dataset type

Track	Model	Dataset	Precision	Recall	Macro F1	Hamming Loss	MR
Track 1	BETO	track1 with emoji	0.9980	0.9991	0.9985		
	BETO	track1 without emoji	0.9984	0.9994	0.9989		
	Robertuito	track1 robertuito	**0.9992**	**0.9998**	**0.9995**		
Track 2	BETO	track2 with emoji	0.8292	**0.8326**	**0.8309**	0.0023	0.9860
	BETO	track2 without emoji	**0.8301**	0.8303	0.8302	**0.0021**	**0.9872**
	TwHINBERT	track2 with emoji	0.8044	0.8258	0.8143	0.0067	0.9617
	TwHINBERT	track2 without emoji	0.8161	**0.8326**	0.8242	0.0050	0.9698
	Robertuito	track2 robertuito	0.7172	0.8303	0.7664	0.0181	0.8932
	BERTIN	track2 with emoji	0.7898	0.8058	0.7972	0.0114	0.9443
	BERTIN	track2 without emoji	0.8103	0.8263	0.8181	0.0061	0.9663
Track3	BETO	track3 preprocessed	0.5085	0.5027	0.4975		
	Alberti	track3 preprocessed	0.4667	0.5000	0.4828		
	MarIA	track3 preprocessed	**0.6689**	**0.6330**	**0.6482**		

5. Results

The organizers used Macro F1 scores as the primary metric to assess participants' performance. To handle submissions and obtain scores for the test dataset, restrictions allowed only 10 distinct results to be submitted. For this work, the top 2 models for each track were selected for submission. The official

results achieved by the top 5 participants for each track are shown in Table 6. Our team secured the 4th, the 3rd, and the 9th positions in each respective track based on the official results. The top models trained during the evaluation step also delivered the best results in the test set. This indicated that other approaches developed in this work were unlikely to improve the final results.

Table 6
Official results per track. Our team achieved 4th, 3rd, and 9th place per each track. The best models are described after the name of our team. For track 3, some participants were omitted.

Track	Participant	Macro F1	Place
Track 1	verbanex	0.9143	1
	atoro491	0.9143	1
	rogerd97	0.9143	1
	CANTeam	0.8775	2
	i2chuelva	0.8764	3
	DSVS (Robertuito)	**0.8713**	**4**
	metztli	0.9143	5
Track 2	CANTeam	0.9730	1
	homomex	0.9487	2
	DSVS (TwHINBERT)	**0.9435**	**3**
	verbanex	0.9393	4
	ajhglez99	0.9345	5
	jmadera	0.9345	5
Track 3	ajhglez99	0.5762	1
	jmadera	0.5762	1
	verbanex	0.5683	2
	metztli	0.5667	3
	jcmqcu	0.5575	4
	carlos31	0.5484	5
	DSVS (BETO)	**0.4864**	**9**

6. Conclusions

This study documents our participation in the HOMO-MEX shared task, focusing on Hate Speech detection in Mexican-Spanish texts. Our engagement in this task resulted in competitive outcomes, particularly notable in tracks 1 and 2. However, in track 3, our performance was hampered by time constraints, leading to the submission of a suboptimal model (BETO and Alberti) and potentially limiting our ability to improve results further. Class weights were implemented based on the balance of the classes per dataset to tackle data imbalance. Nevertheless, the exploration of data augmentation techniques could have provided an additional option for improvement.

In our analysis, we noticed the importance of emoji characters in detecting hate speech in social media posts, especially in tracks 1 and 2. The handling of emoji symbols during the data cleansing process was demonstrated to impact the performance of the models.

Overall, our findings highlighted the complexities involved in detecting hate speech in diverse linguistic contexts. While our participation yielded competitive outcomes, there remains ample room for further exploration and refinement of methodologies to enhance detection accuracy and robustness in real-world applications.

Acknowledgments

This work was done with partial support from the Mexican Government through Consejo Nacional de Humanidades Ciencias y Tecnologías (CONAHCYT) and Instituto Politécnico Nacional (IPN).

References

[1] I. World, Ilga world: International lesbian, gay, bisexual, trans and intersex association, https://ilga.org/, 2024. Accessed: 2024-05-28.

[2] K. Heiden-Rootes, A. Wiegand, D. Bono, Sexual minority adults: A national survey on depression, religious fundamentalism, parent relationship quality & acceptance, Journal of Marital and Family Therapy 45 (2019) 106–119. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jmft.12323. doi:https://doi.org/10.1111/jmft.12323. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jmft.12323.

[3] S. D. Snapp, R. J. Watson, S. T. Russell, R. M. Diaz, C. Ryan, Social support networks for lgbt young adults: Low cost strategies for positive adjustment, Family Relations 64 (2015) 420–430. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/fare.12124. doi:https://doi.org/10.1111/fare.12124. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/fare.12124.

[4] R. Baiocco, L. Fontanesi, F. Santamaria, S. Ioverno, B. Marasco, E. Baumgartner, B. L. B. Willoughby, F. Laghi, Negative parental responses to coming out and family functioning in a sample of lesbian and gay young adults, Journal of Child and Family Studies 24 (2015) 1490–1500. URL: https://doi.org/10.1007/s10826-014-9954-z. doi:10.1007/s10826-014-9954-z.

[5] B. L. Beagan, B. Hattie, Religion, spirituality, and lgbtq identity integration, Journal of LGBT Issues in Counseling 9 (2015) 92–117. URL: https://doi.org/10.1080/15538605.2015.1029204. doi:10.1080/15538605.2015.1029204. arXiv:https://doi.org/10.1080/15538605.2015.1029204.

[6] W. Institute, The economic cost of lgbt stigma and discrimination in south africa, 2019. URL: https://williamsinstitute.law.ucla.edu/events/cost-stigma-s-africa-webinar/.

[7] E. Malkin, A. Ahmed, Mexican president moves to legalize gay marriage nationwide, The New York Times (2016). URL: http://www.nytimes.com/2016/05/18/world/americas/mexico-gay-marriage.html?_r=0, accessed May 01, 2024.

[8] M. Lipka, Same-sex marriage makes some legal gains in latin america, 2015. URL: http://www.pewresearch.org/fact-tank/2015/06/25/same-sex-marriage-makes-some-legal-gains-in-latin-america.

[9] H. R. Campaign, Addressing anti-transgender violence: exploring realities, challenges, and solutions for policy makers and community advocates, 2015. URL: http://www.hrc.org/resources/addressing-antitransgender-violence-exploring-realities-challengesand-sol.

[10] M. Santos, In the shadows: the difficulties of implementing current immigration policies in adjudicating gender-diverse asylum cases in immigration courts, 2012. URL: http://www.hkslgbtq.com/in-the-shadowsthe-difficulties-of-implementing-current-immigrationpolicies-in-adjudicating\-gender-diverse-asylum-cases-inimmigration-courts.

[11] REDLACTRANS, The night is another country: impunity and violence against transgender women human rights defenders in latin america, 2012. URL: http://www.aidsalliance.org/assets/000/000/405/90623-Impunity-and-violence-against-transgenderwomen-human-rights-defenders-in-Latin-\America_original.pdf?1405586435.

[12] 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).

[13] 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.

[14] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).

[15] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del

Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[16] J. de la Rosa, Á. P. Pozo, S. Ros, E. González-Blanco, Alberti, a multilingual domain specific language model for poetry analysis, arXiv preprint arXiv:2307.01387 (2023).

[17] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[18] 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.

[19] X. Zhang, Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, A. El-Kishky, Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations, arXiv preprint arXiv:2209.07562 (2022).

[20] 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).

[21] 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.

[22] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[23] C. F. Rosauro, M. Cuadros, Hate speech detection against the mexican spanish lgbtq+ community using bert-based transformers (2023).

[24] 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).

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

[26] 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: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), 2023.

[27] 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: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), 2023.