LIDOMA at HOMO-MEX2023@IberLEF: Hate Speech Detection Towards the Mexican Spanish-Speaking LGBT+ Population. The Importance of Preprocessing Before Using BERT-Based Models

Moein Shahiki-Tash, Jesús Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov and Alexander Gelbukh

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico city, Mexico
Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC IPN), Juan de Dios Bátiz Av., Gustavo A. Madero, 07738 Ciudad de México, México.

Abstract

Hate speech targeting LGBT+ individuals poses a deeply ingrained problem with wide-ranging consequences, encompassing substance abuse disorders and discrimination. These specific concerns are particularly amplified in Mexico. In this paper, we present our submission on the first track of the HOMO-MEX: Hate Speech Detection towards the Mexican Spanish-Speaking LGBT+ Population. We explore the dataset and we employ transformer architectures, who have demonstrated significant efficacy in similar sentiment analysis tasks. Specifically, we utilize BERT-based models and we show the importance of preprocessing by reaching the last place in the competition with a Macro F1 score of 0.73. The source code to reproduce our results can be found at https://github.com/moeintash72

Keywords

BERT-based models, Hate Speech Detection, LGBT+phobia, Natural Language Processing, Preprocessing, CEUR-WS

1. Introduction

LGBT+phobia, defined as any type of discrimination based on sexual preferences and/or gender identities, remains a significant and far-reaching issue with profound implications. Members of the LGBT+ community are particularly vulnerable to substance abuse disorders, disproportionate mental health challenges, discrimination in labor markets, and limited access to education and healthcare services. These challenges are further amplified in Mexico, where

IberLEF 2023, September 2023, Jaén, Spain

✉ mshahikit2022@cic.ipn.mx (M. Shahiki-Tash); jarmentas2022@cic.ipn.mx (J. Armenta-Segura); zahraahani97100@gmail.com (Z. Ahani); kolesnikova@cic.ipn.mx (O. Kolesnikova); sidorov@cic.ipn.mx (G. Sidorov); gelbukh@cic.ipn.mx (A. Gelbukh)
🌐 https://github.com/moeintash72 (M. Shahiki-Tash); https://sites.google.com/view/jesusarmentasegura (J. Armenta-Segura); https://www.researchgate.net/profile/Olga-Kolesnikova-2 (O. Kolesnikova); https://www.cic.ipn.mx/~sidorov (G. Sidorov); http://www.gelbukh.com (A. Gelbukh)
iD 0009-0003-5767-0566 (M. Shahiki-Tash); 0009-0003-8729-7096 (J. Armenta-Segura); 0000-0002-1307-1647 (O. Kolesnikova); 0000-0003-3901-3522 (G. Sidorov); 0000-0001-7845-9039 (A. Gelbukh)

CEUR Workshop Proceedings (CEUR-WS.org)

substance abuse disorders are highly prevalent even beyond the LGBT+ community [1].

The HOMO-MEX: Hate Speech Detection towards the Mexican Spanish-Speaking LGBT+ Population [2] includes a tracks aimed at addressing the issue, by detecting whether a Mexican tweet contains LGBT+phobia or not, differentiating between the tweets who address the topics from the ones who does not. Hate speech detection is a challenging task due to the complex interplay of linguistic factors and nuanced emotions involved in it [3]. To address this challenge, the use of transformers has been instrumental, with promising results in homophobia detection and other related text processing tasks [4, 5, 6, 7, 8].

Transformers [9] have revolutionized natural language processing tasks with the introduction of the self-attention mechanism. This mechanism empowers the model to capture global dependencies and contextual relationships within a sequence, rendering transformers highly adept at handling tasks such as sentiment analysis and, more notably, hate speech, as previously mentioned. While this may seem promising on paper, it is crucial to preprocess the text to optimize the efficacy of attention mechanisms. By removing *lexical noise*, such as stopwords, and eliminating *linguistic noise* through lemmatization, self-attention mechanisms can focus exclusively on pragmatic and metalinguistic features, which are directly associated with the text representation of sentiments. As a dramatic proof of the importance of preprocessing, our submission omitted it and got the last place on the competition.

In this paper, we describe our system to address this shared task. We obtained a Macro F1 score of 0.73. The structure of this paper is as follows: in Section 2, we describe some state of the art works on hate speech detection and LGBT+phobia detection. In Section 3, we detail our methodology. In Section 4, we provide a description for the dataset. Additionally, we outline our experimental workflow. In Section 5, we discuss the results of our experiments. Finally, in Section 6, we conclude the paper.

2. Related Work

Hate speech has been a prominent area in the field of Natural Language Processing for quite some time [10, 11], with significant attention being drawn to it at least since 1997 [12], where a system called *Smokey* was developed to detect *abusive messages* on internet.

In [13], the authors proposed that hate speech could be understood in terms of its lexicon and how that is used, resembling the task of word sense disambiguation. However, their findings indicated that this hypothesis might not hold when dealing with incomplete datasets. For instance, the word *jew* was exclusively present in anti-semitic speech of their dataset, causing their methods to associate *jew* as a solely anti-Jewish feature, regardless of its intended sense. In [14], the authors made an extensive classification of hate speech targets, which allowed them to focus in sentence structure more than in lexical features. As a consequence, they were able to detect racial speech regardless racial slur such as *nigga*, which is also used among black people in a non-racist way. However, they also accepted that their approach was not extensive enough to tackle the task.

Particularly, hate speech against LGBT+ population has also received a singular focus, as in [15], where a structuralist approach were employed to manually characterize linguistic features in african homophobic speech, or in [16], where those features were addressed from a most automatical and computational point of view.

Another important task related to hate speech detection is the identification of supportive speech, which was addressed in the same edition of IberLEF through a shared task called Hope Speech [17]. In that shared task, we made two submissions [18, 19] which achieved third and fourth place, respectively.

Focusing on the computational point of view, several tools and techniques from Natural Language Processing have been used to address the task, such as TF/IDF with bigrams [20, 21, 22] or the zero-shot learning paradigm [23], who achieved several first places on the LT-EDI-ACL2022 homophobia/transphobia speech detection contest [24], until the current state-of-the-art transformers-based approaches [16, 25, 26, 27]. As long as we understood, all these works preprocessed the dataset before training the model.

3. Methodology

Our main goal was to develop a method capable to focus in the relevant features related with LGBT+phobia, based on transformers. Keeping that on mind, we first converted all labels into numeric values, as depicted in table 1.

Table 1
Class labeling for preprocessing

Class	Original label	Numerical label
non LGBT+phobic tweet (and address the issue)	NP	0
LGBT+phobic tweet	P	1
non LGBT+phobic tweet (does not address the issue)	NA	2

After that, we converted the dataset into a Hugging Face arrow dataset and we tokenized all tweets using the *bert-base-cased* model [28], which allow us to retrieve special tokens and their corresponding IDs. The next step was to start the classification task, for which we loaded the Auto-Model-For-Sequence-Classification class from the Hugging Face library, in order to use a BERT model. Finally, we used the best trained model to make predictions on the test dataset.

4. Experimental Setup

4.1. Data

The task involves analyzing a large dataset of $7,000$ Mexican tweets gathered from 2012 to 2021. As showed in Table 1, they are labelled as P for LGBT+phobic tweets, NP for non-LGBT+phobic tweets who address the issue, and NA for non-related tweets. See Table 2 for the precise number of tweets for each class.

Table 2
Statistics of the dataset

Class	Ammount of tweets
non LGBT+phobic tweet (and address the issue)	$4,360$
LGBT+phobic tweet	$1,778$
non LGBT+phobic tweet (does not address the issue)	862

In Table 3 there are six examples from the dataset. Some of the tweets labelled as LGBT+phobics consist of authors not showing such behavior, but describing someone who does. For instance, in Example 0, the author describes what happened when he tried to hit on a lesbophobic girl while, in Example 2, the author is a journalist exposing a lesbophobic tycoon. Moreover, some non-LGBT+phobic tweets were wrong labelled, as Example 1140, who is directly homophobic given the Mexican idiosyncrasy.

Table 3
Six selected examples from the training dataset

Tweet id	Tweet in spanish	Label	English Translation
0	Me quise ligar a una chava ayer y no me pelo, le pregunte si era lesbiana y me dio una cachetada	P	Yesterday, I tried to hit on a girl but she ignored me. I asked her if she was lesbian, and she slapped me.
2	Magnate ofrece 130 mdd al hombre que conquiste a su hija lesbiana	P	Tycoon offers 130 mdd to the man who manages to seduce his lesbian daughter.
81	Ser tortillero me hace lesbiana?	P	Being tortilla-maker makes me lesbian?
1140	ese wero afilador de puñales!!! jajaja	NP	hey you, blond fag-maker!!! hahaha
1165	Jjajajajajaja ayñs perdon por no estar en la onda, pinche jota!!	NP	ROLF, so sorry for not be into the current trend, fucking faggot!!
508	FelixDice: marica burra y gallina clueca	NA	FelixSays: fag donkey and broody hen.

4.2. Experimental workflow

Our experimental setup is as follow:

- We prepared the data by converting the labels into numerical values as depicted in Table 1
- Then, we tokenized with the Hugging Face AutoTokenizer, truncated to a maximum length of 32 tokens, and padded to ensure consistent input size.
- For the classification task, as hinted in Section 3, we loaded the pre-trained BERT model for sequence classification (AutoModelForSequenceClassification) with the specified number of labels (three, in this case).
- Finally, we trained the model with the *Trainer* class from the Hugging Face library.

Training arguments were configured, including the output directory, logging settings, batch size, learning rate, and number of epochs. This class also included backpropagation and optimization.

Once with a model trained in the test dataset, we use it for make predictions on the test dataset. The predicted labels are obtained from the predicted probabilities using argmax. We saved the results in the text file *result.txt* with tab-separated values (TSV) format.

5. Results

After a four-epoch training, we achieved a Macro F1 score of 0.73. For the results, two main factors were crucial: in the first place, the lack of a preprocessing before the tokenization with the Huggingface AutoTokenizer, and in second place the quality of the dataset labeling. As showed in Section 4, Table 3, several LGBT+phobic instances described someone with that behavior, but distinct of the author. Carving on the dataset, we also find non-LGBT+phobic instances with the same quirk (see two examples in Table 4), who might led to ambiguity to the trained model.

Table 4
Two examples from the training dataset of non-LGBT+phobic tweets whose author describes a LGBT+phobic behavior

Tweet id	Tweet in spanish	English Translation
2788	Wow, luego que las TERFas se aventajaran en el Reino Unido ahora ya lograron que el sistema médico destransicione a la gente joven trans	Wow, after TERF feminists gained power in the UK, they obligue the medical system to untransition young trans people.
2795	Para que luego a una como mujer trans no la dejen entrar a estas cosas o se la hagan de jamón. ¡Gracias, nerdos tóxicos!	To provoke them to not allow you to enter in this stuff, due you are a trans woman. Thank you!, toxic nerds!

Another opportunity of enhancemnt can be found in the non-LGBT+phobic labeling. For instante, in Examples 1140 and 1165, we find adjectives for LGBT+ people used as an insult, but the tweet was labelled as non-LGBT+phobic. Example 508 is a LGBT+phobic tweet, labelled as non-related. From the three labels, we empirically suspect that the best was the non-related, since we struggle to find a wrong example. However, in the other two examples, a quick search of key words was enough to find several opportunities of enhancement, which lead us to suspect that the mentioned quirks are statistical significative enough to bias the state-of-the-art BERT models.

6. Conclusions

In this paper, we presented our approach for addressing the first track of the HOMO-MEX 2023 shared task of hate speech detection towards the Mexican Spanish-speaking LGBT+ population. We leveraged the power of transformers, specifically BERT models, which have proven effective in hate speech tasks.

Our methodology involved the conversion of labels into numerical values, and tokenizing the tweets using the bert-base-cased model, without a text-preprocessing. We employed the AutoModelForSequenceClassification class from the Hugging Face library to train the BERT model and make predictions on the test dataset.

Although the lack of preprocessing influenced a lot the poor results, we also find several opportunities for improvement in the dataset, with instances of mislabeling and ambiguity. These discrepancies might introduce biases and affect the efficacy of state-of-the-art BERT models.

The findings highlight the complexities involved in detecting LGBT+phobic speech and call for continued research and development in addressing this critical issue. Future work should focus on enhancing the classifications and embeddings, maybe by considering tools with less use in this tasks like the zero-shot learning paradigm. Also, we want to explore novel approaches to improve the detection and mitigation of LGBT+phobia in online spaces.

7. Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20232138, 20232080, 20231567 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

[1] Secretaría de Salud, Encuesta Nacional de Consumo de Drogas 2016-2017, Survey, 2017. URL: https://www.gob.mx/cms/uploads/attachment/file/234856/CONSUMO_DE_DROGAS.pdf, mexico.

[2] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S.-T. Andersen, S. Ojeda-Trueba, Overview of HOMO-MEX at Iberlef 2023: HOMO-MEX: Hate speech detection in Online Messages directed tOwards the MEXican spanish speaking LGBTQ+ population, Procesamiento del lenguaje natural 71 (2023).

[3] F. Balouchzahi, H. Shashirekha, Las for hasoc-learning approaches for hate speech and offensive content identification, 2020.

[4] N. Ghanghor, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, S. Thavareesan, B. R. Chakravarthi, IIITK@LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil , Malayalam and English, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Kyiv, 2021, pp. 197–203. URL: https://aclanthology.org/2021.ltedi-1.30.

[5] Z. M. Farooqi, S. Ghosh, R. R. Shah, Leveraging transformers for hate speech detection in conversational code-mixed tweets, 2021. arXiv:2112.09986.

[6] T. Ranasinghe, S. Gupte, M. Zampieri, I. Nwogu, Wlv-rit at hasoc-dravidian-codemix-fire2020: Offensive language identification in code-switched youtube comments, 2020. arXiv:2011.00559.

[7] M. Gemeda Yigezu, A. Lambebo Tonja, O. Kolesnikova, M. Shahiki Tash, G. Sidorov, A. Gelbukh, Word level language identification in code-mixed Kannada-English texts using deep learning approach, in: Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 29–33. URL: https://aclanthology.org/2022.icon-wlli.6.

[8] A. L. Tonja, M. G. Yigezu, O. Kolesnikova, M. S. Tash, G. Sidorov, A. Gelbuk, Transformer-based model for word level language identification in code-mixed kannada-english texts, 2022. arXiv:2211.14459.

[9] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017.

[10] 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.

[11] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1 – 30.

[12] E. Spertus, Smokey: Automatic recognition of hostile messages, in: AAAI/IAAI, 1997.

[13] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the Second Workshop on Language in Social Media, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 19–26. URL: https://aclanthology.org/W12-2103.

[14] L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, I. Weber, Analyzing the targets of hate in online social media, CoRR abs/1603.07709 (2016). URL: http://arxiv.org/abs/1603.07709.

arXiv:1603.07709.

[15] V. Reddy, Perverts and sodomites: homophobia as hate speech in africa, Southern African Linguistics and Applied Language Studies 20 (2002) 163 – 175.

[16] I. S. Upadhyay, K. A. Srivatsa, R. Mamidi, Sammaan@LT-EDI-ACL2022: Ensembled transformers against homophobia and transphobia, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 270–275. URL: https://aclanthology.org/2022.ltedi-1.39. doi:10.18653/v1/2022.ltedi-1.39.

[17] Jiménez-Zafra, Salud María and García-Cumbreras, Miguel Ángel and García-Baena, Daniel and García-Díaz, José Antonio and Raja Chakravarthi, Bharathi and Valencia-García, Rafael and Ureña-López, L. Alfonso, Overview of HOPE2023@IberLEF: Multilingual Hope Speech Detection, Procesamiento del Lenguaje Natural 71 (2023).

[18] Z. Ahani, G. Sidorov, O. Kolesnikova, A. Gelbukh, Hope speech detection from text using tf-idf features and machine learning algorithms, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF), Jaén, Spain, 2023.

[19] M. Shahiki-Tash, J. Armenta-Segura, O. Kolesnikova, G. Sidorov, A. Gelbukh, LIDOMA at HOPE2023@IberLEF: Hope speech detection using lexical features and convolutional neural networks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF), Jaén, Spain, 2023.

[20] N. Ashraf, M. Taha, A. Abd Elfattah, H. Nayel, NAYEL @LT-EDI-ACL2022: Homophobia/-transphobia detection for equality, diversity, and inclusion using SVM, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 287–290. URL: https://aclanthology.org/2022.ltedi-1.42. doi:10.18653/v1/2022.ltedi-1.42.

[21] M. Shahiki Tash, Z. Ahani, A. Tonja, M. Gemeda, N. Hussain, O. Kolesnikova, Word level language identification in code-mixed Kannada-English texts using traditional machine learning algorithms, in: Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 25–28. URL: https://aclanthology.org/2022.icon-wlli.5.

[22] F. Balouchzahi, H. L. Shashirekha, G. Sidorov, Hssd: Hate speech spreader detection using n-grams and voting classifier, in: CEUR Workshop Proceedings, volume 2936, 2021, pp. 1829–1836. URL: www.scopus.com, cited By :6.

[23] M. Singh, P. Motlicek, IDIAP submission@LT-EDI-ACL2022: Homophobia/transphobia detection in social media comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 356–361. URL: https://aclanthology.org/2022.ltedi-1.55. doi:10.18653/v1/2022.ltedi-1.55.

[24] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. McCrae, P. Buitelaar, P. Kumaresan, R. Ponnusamy, Overview of the shared task on homophobia and transphobia detection in social media comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 369–377. URL: https://aclanthology.org/2022.ltedi-1.57. doi:10.18653/v1/2022.ltedi-1.57.

[25] D. Nozza, Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 258–264. URL: https://aclanthology.org/2022.ltedi-1.37. doi:10.18653/v1/2022.ltedi-1.37.

[26] V. Bhandari, P. Goyal, bitsa_nlp@LT-EDI-ACL2022: Leveraging pretrained language models for detecting homophobia and transphobia in social media comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 149–154. URL: https://aclanthology.org/2022.ltedi-1.18. doi:10.18653/v1/2022.ltedi-1.18.

[27] A. Maimaitituoheti, ABLIMET @LT-EDI-ACL2022: A roberta based approach for homophobia/transphobia detection in social media, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 155–160. URL: https://aclanthology.org/2022.ltedi-1.19. doi:10.18653/v1/2022.ltedi-1.19.

[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, ArXiv abs/1810.04805 (2019).

[29] J. Armenta-Segura, G. Sidorov, A baseline for anime success prediction, based on synopsis, in: Congreso Mexicano de Inteligencia Artificial de la Sociedad Mexicana de Inteligencia Artificial COMIA-MICAI 2023, Zapopan, Jalisco, 2023.