Impact of Text Preprocessing and Feature Selection on Hate Speech Detection in Online Messages Towards the LGBTQ+ Community in Mexico

Cesar Macias*,†, Miguel Soto†, Tania Alcántara† and Hiram Calvo†

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

Abstract

The prevalence of online hate speech targeting the LGBTQ+ community poses a significant challenge in maintaining a safe and inclusive digital environment. This paper deals with the importance of addressing this issue by proposing methods for detecting this offensive messages towards this community population in Mexican Spanish. The study explores a considerable variety of approaches to solve the task with classical machine learning algorithms and with different approaches for feature extraction. Additionally, text preprocessing techniques specific to Twitter data, and word embeddings are employed to enhance the performance of the models. Through experimentation and comparative analysis, we assess the effectiveness of these methods in identifying and classifying offensive messages. The findings of this research contribute to the development of robust tools for identifying and mitigating online hate speech, ultimately fostering a more inclusive and tolerant digital space for the LGBTQ+ community.

Keywords

Hate speech, gender identities, natural language processing, machine learning.

1. Introduction

Twitter is an online social networking platform that enables users to send and read short messages called tweets. Tweets are messages of up to 280 characters that can contain text, links, images, videos, and other multimedia content. Twitter users can follow others and see their tweets in their timeline, allowing them to keep up to date on the news, updates, thoughts, and opinions. Users can also interact with other users tweets through replies, retweets (sharing someone else's tweet on your own timeline), and liking tweets. Due to the ease of access to the platform, and the information it contains, information flows all the time, and we can not only find positive messages or information, but also hate messages.

Given the anonymity that social networks provide to users, they have become a great source of offensive language. It has been shown that there is a strong correlation between anonymity

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

†These authors contributed equally.

cmaciass2021@cic.ipn.mx (C. Macias); msotoh2021@cic.ipn.mx (M. Soto); talcantaram2020@cic.ipn.mx (T. Alcántara); hcalvo@cic.ipn.mx (H. Calvo)

https://github.com/MACI-dev-96 (C. Macias); https://github.com/mashd3v (M. Soto); http://hiramcalvo.com/ (H. Calvo)

0000-0003-2836-2102 (H. Calvo)

and hate speech. In particular, hate speech about race and sexual orientation is more likely to be posted anonymously compared to other categories of hate speech [1]. This aspect also applies to hate speech because the Internet has a democratic character and offers the opportunity for all people to engage in hate speech [2]. Previous research has identified that, compared to heterosexuals, LGBTQ+ people are at higher risk of developing poor mental health, problems related to excessive alcohol consumption and smoking [3]. In addition, transgender people are more likely to develop depressive symptoms and experience stress compared to non-transgender people in the LGBTQ+ community [4]; also, transgender people are more likely to experience hate crimes [5].

The discrimination against LGBTQ+ people begin at home. Alone, 92% of adolescents with these preferences had to hide their sexual orientation until adulthood. This discrimination suffered at home, is transferred to virtual environments, such as Twitter, where there is a more constant attack by the way they dress, speak and even how they write on the platform.

Within the framework of the IberLEF (Iberian Languages Evaluation Forum) 2023 congress, the Grupo de Ingeniería Lingüística at the Universidad Nacional Autónoma de México, proposed the task HOMO-MEX: Hate speech detection in Online Messages directed tOwards the MEXican spanish speaking LGBTQ+ population [6], with the aim of identifying texts that exhibit phobic content towards the LGBTQ+ community.

The solution proposed by our team is described in the following sections. This work is structured as follows. In Section 2 a brief description of the related works in hate speech detection is given. In Section 3, the description of the datasets, the preprocessing, and feature extraction methods accompanied by the classification models implemented is given. Then, in Section 4 we describe the setup for the experiments and the results obtained after the submission of our predictions. Finally, in Section 5 we discuss our results and some ideas about how to improve them for future work.

2. State of the Art

Since the early detection of hate speech in social networks has been considered a major problem to be solved in recent years, numerous contributions have been made both in terms of corpus and approaches to solve this task automatically with algorithms and artificial intelligence (AI) models. This section is structured as follows: (1) Some works with AI approaches for hate speech detection will be reviewed. (2) Some works with AI approaches for hate speech detection specific to the LGBT+ community will be reviewed. (3) The datasets available to deal with this task will be mentioned.

2.1. Hate speech detection

In 2004, Greevy [7], carried out an approach for the detection of racism on the Internet. To achieve this, he made use of SVM, and used as features: BoW, n-grams with 2- and 3-word sequences and a reduced version of POS.
By 2015, Burnap and Williams [8], used BoW, n-grams with two-word sequences, and typed dependencies as feature extraction. Later in 2016, Burnap and colleagues [9], would implement

an approach with two machine learning models: linear SVMs and decision trees. For them, they considered sequences of up to 5 n-grams, lemmas, and offensive language terms and phrases.
In 2017, Davidson et al. [10], first proposed to use the logistic regression classification model with a self-created dataset. However, they later tried a variety of other models such as: Naive Bayes, decision trees, random forests, and linear SVMs. Later that year, Vignia et al. [11] proposed the use of two different classifiers. The first one based on SVM and the second one was a long-short term memory (LSTM) network where they seek to capture long-range dependencies in tweets that may play an important role.
More recently, in 2020, Ozler et al. [12], suggest the use of a fine-tuning of the BERT model [13] for multi-domain, multi-class non-civil language prediction. In the same year, similarly, Mozafari et al. [14] and Khan et al. [15] present BERT-based architectures to deal with the task of hate speech detection.

2.2. Hate speech towards LGBTQ+ community detection

Speaking specifically about the detection of hate speech against the LGBT+ community, there are not many works on the subject currently. However, recently in 2021 Dias et al. [16], fights hate speech through the implementation of an artificial intelligence model to moderate the content and risks of LGBT+ voices online. To achieve this, they used Perspective, an AI technology developed by Jigsaw (formerly Google Ideas). With this tool, they measure the perceived levels of toxicity of text-based content.
By 2022, Hartvigsen et al. [17] propose a comprehensive computer-generated dataset for detecting implicit hate speech and adversarial hate speech called *ToxicGen*. In this dataset they focused on minorities including the LGBTQ+ community, to test the effectiveness of this dataset they made use of HateBERT [18] and ToxDectRoBERTa [19].

2.3. Available datasets

A summary of some of the datasets that are available to date for dealing with hate speech can be found in Table 1.

3. Methodology

In this section, we briefly describe the datasets provided for the competition, and we provide a more detailed description of the methods used to preprocess the text, how we extracted the features from the plain text to vectorize it and the models used to perform the classification.

3.1. Dataset description

For both tracks, just the training dataset with tags were given to the competitors (test dataset only contains the documents but no tag is contained).

3.1.1. Track 1: Hate speech detection track (Multi-class)

For this track, the training dataset contained documents tagged as:

Table 1
Available hate speech datasets.

Name	Data source	Labels	Language	Reference
Waseem	Twitter	Racism, Sexism	English	[20]
Davidson	Twitter	Hate speech, Offensive	English	[10]
Founta	Twitter	Hate speech, Offensive	English	[21]
HatEval	Twitter	Hateful	English	[22]
Kaggle	Wikipedia	Toxic, Severe Toxic, Obsceene, Theat, Insult, Identity Hate	English	Kaggle
AMI	Twitter	Misogynous	English	[23]
Warner	Yahoo! American Jewish Congress	Anti-semitic, Anti-black, Anti-asian, Anti-woman, Anti-muslim, Anti-immigrant, Other-hate	English	[24]
TOXIGEN	LLMs prompting	Black, Asian, Native Am, Latino, Jewish, Muslim, Chinese, Mexican, Middle Eastern, LGBTQ+, Women, Mental Dis, Physical Dis	English	[17]
HOMO-MEX	Twitter	LGBT+phobic (P), not LGBT+phobic (NP), not LGBT+related (NA), Lesbophobia (L), Gayphobia (G), Biphobia (B), Transphobia (T), other LGBT+phobia (O).	Spanish	[6]

- LGBT+ phobic (P)
- not LGBT+ phobic (NP)
- not LGBT+ related (NA)

This dataset contained a total of $7,000$ tweets written in Mexican Spanish.

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

The training dataset provided for this track has a total of 862 tweets written in Mexican Spanish. The documents contained in the dataset were multi-labeled, with at least one of the following options:

- Lesbophobia (L)
- Gayphobia (G)
- Biphobia (B)
- Transphobia (T)
- other LGBT+ phobia (O)

3.2. Preprocessing

For both track 1 and track 2, the preprocessing methods applied were the same. Those methods are described below.

- HTML entities: the HTML entities like were removed.
- Line breaks: all the line breaks were removed to obtain a single plain text.
- Twitter entities: Twitter entities like #Hashtag, @User, or URLs were replaced by a special tag for each of them like _HASHTAG_, _USER_, _URL_.
- Lowercase: all the uppercase letters turned to lowercase.
- Apostrophes: all of them were removed.
- Punctuation: the characters used to punctuate the text (e.g., . : , ;) were removed.
- Repeated characters: when a word contains a character that repeats more than twice, the characters were trimmed into two repetitions (e.g., *buenaaaaaas* $\rightarrow$ *buenaas*)
- Alphanumeric words: the words composed by alphabetic and numerical characters like in leet speaking (e.g., *P3nd3j0*) were removed.

3.3. Feature extraction

3.3.1. Bag of words

This model is a simplified representation of a text, where the bag represents the set of all words contained in a document collection. A single document is represented as a vector, where each dimension represents a word from the vocabulary obtained from the document collection and how many times the word appeared in a single document (TF).

3.3.2. TF·IDF

The TF·IDF method makes use of the term frequency (TF), and the inverse document frequency (IDF). For the TF calculation, a bag of words is generated with the vocabulary of the set of all documents that are to be analyzed, then the total number of occurrences of the word is obtained. IDF is calculated as the logarithm of the quotient of the number of documents in which the analyzed word appears and the number of documents in which it appears in the analyzed set. Finally, the product between TF and IDF is performed for each of the words in the analyzed text, generating a vector of characteristics.

3.4. Models

3.4.1. Track 1: Hate speech detection

Linear Support Vector Machine (LSVM). SVM is a classification algorithm that receives input data, then those inputs are mapped to an n-dimentional space. Once the inputs are mapped, the machine uses the support vectors in the boundaries of a class "cluster" to generate a hyperplane able to accurately separate the data into the training classes. The distance between the support vectors and the hyperplane is known as margin. The goal of the SVM is to maximize the margin. If the space is not adequate to optimize the hyperplane, a function (kernel) is

used to perform a spatial transformation to the data, so the machine is able to operate in high dimensional spaces. Linear Support Vector Machines function like a SVM, the only difference is that the SVM is going to use the functions contained in the linear kernel set to find the optmimal solution for the classification problem.

Bagging Classifier. To give solution to the classification problem of track 2, we used a model ensemble of Linear Support Vector Machines (LSVMs). To implement the ensemble, we selected the `BaggingClassifier` from the `Scikit-learn` library [25]. The classifier fits base classifiers on subsets obtained randomly from the original training set, to make predictions, the classifier obtains the classification prediction of each of the base classifiers and make the final prediction by averaging them or by voting them.

3.4.2. Track 2: Fine-grained hate speech detection

Decision Trees (DT). This model uses the symbolist paradigm. To learn the knowledge contained in the training dataset, the decision tree infer simple decision rules. The DT are easy to undestand and interpret and is able to handle multi-output problems [25]. This is one of the reasons to select this classifier.

Multi Output Classifier (MO). To overcome the multi-label nature of this track, we decided to implement a `MultiOutputClassifier` from `scikit-learn` library [25]. This classifier fits one base binary classifier for each of the labels.

4. Experiments and results

In this section, we explain the setup for the experiments performed for the official final results obtained for track 1 and 2 respectively.

4.1. Track 1: Hate speech detection track

The setup of the experiments for the first track is described below. To vectorize the data, `TfIdfVectorizer` with the following parameters:

- `ngram_range= (1, 2)`
- `min_df= 3`

The classifier used was the `BaggingClassifier` with `LSVM` estimators ant the following parameters:

- `estimator= LinearSVC`
  - `penalty= 'l2'`
  - `C= 1.0`
  - `random_state= 42`
- `n_estimators= 15`

Table 2
Track 1 submissions. Final prediction scores.

Rank	Team	Score
1	bayesiano98	0.8847
2	carfer	0.8432
3	JoseAGD	0.8421
4	homomex23	0.8390
5	Cordyceps	0.8354
6	moranoroman	0.8325
7	UTB_NLP	0.8202
8	mgraffg	0.8050
9	Mesay	0.7967
10	**cesar_m**	**0.7635**

- `random_state= 42`

The above listing shows the parameters used for the `BaggingClassifier` on first level of the listing, and the parameters used for the base estimator `LinearSVC` on the second level of the listing.

After the submission was made, we obtained a F1-score of 0.7635 and the comparison between our results and the members of the competition are displayed in Table 2.

4.2. Track 2: Fine-grained hate speech detection track

The setup of the experiments for the first track is described below. To vectorize the data, `TfIdfVectorizer` with the following parameters:

- `ngram_range= (1, 2)`
- `min_df= 3`

The classifier used was the `MultiOutputClassifier` with `DecisionTreeClassifier` estimators ant the following parameters:

- `estimator= DecisionTreeClassifier`
  - `criterion= 'gini'`
  - `splitter= 'best'`
  - `max_depth= 15`
  - `random_state= 42`

The above listing shows the parameters used for the `MultiOssutputClassifier` on first level of the listing, and the parameters used for the base estimator `DecisionTreeClassifier` on the second level of the listing.

After the submission was made, we obtained a macro-average F1 of 0.6550 and the comparison between our results and the members of the competition are displayed in Table 3.

Table 3
Track 2 submissions. Final prediction scores

Rank	Team	Score
1	moranoroman	0.6960
2	carfer	0.6847
3	ErikaRivadeneira	0.6834
4	bayesiano98	0.6812
5	Cordyceps	0.6793
6	Mesay	0.6733
7	homomex23	0.6703
8	JoseAGD	0.6687
9	**cesar_m**	**0.6550**

5. Discussion

On one hand, for task 1, multiple models were evaluated using different data vectorization techniques. The best performing model was the `BaggingClassifier` with LSVM estimators. For data vectorization, TF·IDF with 2-word n-gram sequences was used. Additionally, a minimum word frequency of 3 was set. The results showed that this model achieved a good performance in detecting hate speech messages towards the LGBTQ+ community in Mexican Spanish. On the other hand, for the fine-grained hate speech detection task, different approaches were applied using the `MultiOutputClassifier` with `DecisionTreeClassifier` estimators. For data vectorization, TF·IDF with 2-word n-gram sequences was again used, and the minimum word frequency of 3 was maintained. This model proved to be effective in detecting different levels of hate speech, allowing for a more detailed and accurate classification of offensive messages.

The results obtained in both tasks highlight the importance of using machine learning-based approaches for detecting online hate speech directed towards the LGBTQ+ community in Mexican Spanish. The use of vectorization techniques such as TF·IDF with 2-word n-gram sequences helped capture relevant linguistic features of the messages, enhancing the models' ability to identify hate speech patterns.

Importantly, setting a minimum frequency of occurrence of 3 words helped to filter out infrequent terms and noise in the data, which helped to improve the quality of the predictions.

These results suggest that the combination of appropriate classification models and effective vectorization techniques can be crucial to achieve accurate and effective detection of online hate speech towards the LGBTQ+ community in Mexican Spanish.

Acknowledgments

Aknowledgements to Consejo Nacional de Humanidades Ciencias y Tecnologías (CONAHCYT) and Instituto Politécnico Nacional (IPN).

References

[1] M. Mondal, L. A. Silva, F. Benevenuto, A measurement study of hate speech in social media, in: Proceedings of the 28th ACM conference on hypertext and social media, 2017, pp. 85–94.

[2] A. Brown, What is so special about online (as compared to offline) hate speech?, Ethnicities 18 (2018) 297–326.

[3] K. I. Fredriksen-Goldsen, H.-J. Kim, S. E. Barkan, A. Muraco, C. P. Hoy-Ellis, Health disparities among lesbian, gay, and bisexual older adults: Results from a population-based study, American journal of public health 103 (2013) 1802–1809.

[4] K. I. Fredriksen-Goldsen, L. Cook-Daniels, H.-J. Kim, E. A. Erosheva, C. A. Emlet, C. P. Hoy-Ellis, J. Goldsen, A. Muraco, Physical and mental health of transgender older adults: An at-risk and underserved population, The Gerontologist 54 (2014) 488–500.

[5] M. A. Walters, J. Paterson, R. Brown, L. McDonnell, Hate crimes against trans people: assessing emotions, behaviors, and attitudes toward criminal justice agencies, Journal of interpersonal violence 35 (2020) 4583–4613.

[6] 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).

[7] E. Greevy, Automatic text categorisation of racist webpages., Ph.D. thesis, 2004.

[8] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & Internet 7 (2015) 223–242. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.85. doi:https://doi.org/10.1002/poi3.85. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.85.

[9] P. Burnap, M. Williams, Us and them: identifying cyber hate on twitter across multiple protected characteristics, EPJ Data Science 5 (2016). doi:10.1140/epjds/s13688-016-0072-6.

[10] T. Davidson, D. Warmsley, M. W. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: ICWSM, 2017.

[11] F. D. Vigna, A. Cimino, F. dell'Orletta, M. Petrocchi, M. Tesconi, Hate me, hate me not: Hate speech detection on facebook, in: A. Armando, R. Baldoni, R. Focardi (Eds.), Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017, volume 1816 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 86–95. URL: http://ceur-ws.org/Vol-1816/paper-09.pdf.

[12] K. B. Ozler, K. Kenski, S. Rains, Y. Shmargad, K. Coe, S. Bethard, Fine-tuning for multi-domain and multi-label uncivil language detection, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 28–33. URL: https://aclanthology.org/2020.alw-1.4. doi:10.18653/v1/2020.alw-1.4.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[14] M. Mozafari, R. Farahbakhsh, N. Crespi, Hate speech detection and racial bias mitigation

in social media based on bert model, PLOS ONE 15 (2020) 1–26. URL: https://doi.org/10.1371/journal.pone.0237861. doi:10.1371/journal.pone.0237861.

[15] S. Khan, M. Fazil, V. K. Sejwal, M. A. Alshara, R. M. Alotaibi, A. Kamal, A. R. Baig, Bichat: Bilstm with deep cnn and hierarchical attention for hate speech detection, Journal of King Saud University - Computer and Information Sciences 34 (2022) 4335–4344. URL: https://www.sciencedirect.com/science/article/pii/S1319157822001550. doi:https://doi.org/10.1016/j.jksuci.2022.05.006.

[16] T. Dias Oliva, D. M. Antonialli, A. Gomes, Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online, Sexuality & Culture 25 (2021) 700–732.

[17] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, E. Kamar, Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection, 2022. arXiv:2203.09509.

[18] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, Hatebert: Retraining bert for abusive language detection in english, arXiv preprint arXiv:2010.12472 (2020).

[19] X. Zhou, Challenges in automated debiasing for toxic language detection, University of Washington, 2021.

[20] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013. doi:10.18653/v1/N16-2013.

[21] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, in: Proceedings of the international AAAI conference on web and social media, volume 12, 2018.

[22] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[23] E. Fersini, D. Nozza, P. Rosso, et al., Overview of the evalita 2018 task on automatic misogyny identification (ami), in: EVALITA Evaluation of NLP and Speech Tools for Italian Proceedings of the Final Workshop 12-13 December 2018, Naples, Accademia University Press, 2018.

[24] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the second workshop on language in social media, 2012, pp. 19–26.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.