Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov and Alexander Gelbukh

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), NLP Lab, Mexico City, Mexico.

Abstract

This paper focuses on identifying hate speech directed towards the LGBT+ community. The study involves two tasks, track 1 and track 2, which use a multi-class approach to identify LGBT+phobic content in tweets and detect fine-grained multi-label hate speech indicating different types of LGBT+phobias, respectively. The study employs pre-processing and oversampling techniques to address data imbalance problems. The results show that transformer-based approaches, such as BERT and RoBERTa, are effective in identifying hate speech directed at the LGBT+ community. The experiment performance is evaluated by the macro-average F1 measure. The study highlights the challenges associated with data imbalance, order bias, and limited training data, which can lead to bias in model performance and affect its ability to learn the underlying patterns in the data.

Keywords

Transformer-based, Hate speech, Multi-labeled, Multi-class, BERT, Roberta, Social media

1. Introduction

Recent advances in mobile computing and the internet have led to social media being used to communicate, express ideas, interact with others, and share information. While social media provides a valuable way to communicate easily and efficiently, it also serves as a tool for spreading hate speech online. Internet features often contribute to the misuse of social networks to transmit and spread hate speech [1].

It is not uncommon for someone to take advantage of and misuse social media networks in order to disseminate content that is insulting, abusive, or otherwise detrimental to other users. Every type of online platform on which user-generated information is shown, such as the comment sections of news websites and real-time chat rooms, is now facing a serious challenge in the form of the proliferation of hate speech. In legal and academic literature, speech that conveys hatred towards a person or group is commonly referred to as hate speech [2].

Hate speech detection is one of the essential activities in Natural Language Processing (NLP) fields which is a method used to identify hate speech in various social media platforms.

IberLEF 2023, September 2023, Jaén, Spain

mgemedak2022@cic.ipn.mx (M. G. Yigezu); kolesolga@gmail.com (O. Kolesnikova); sidorov@cic.ipn.mx (G. Sidorov); gelbukh@cic.ipn.mx (A. Gelbukh)

https://mesay-gemeda.github.io/ (M. G. Yigezu); https://www.cic.ipn.mx/~sidorov/ (G. Sidorov); http://www.gelbukh.com/ (A. Gelbukh)

0000-0003-1913-2612 (M. G. Yigezu); 0000-0002-1307-1647 (O. Kolesnikova); 0000-0003-3901-3522 (G. Sidorov); 0000-0001-7845-9039 (A. Gelbukh)

As hate speech has led to discrimination, harassment, and violence against individuals or groups based on their race, religion, gender, sexuality, or other personal characteristics [3, 4], by it, we can take steps to prevent these harms, protect vulnerable populations and get the following benefit:-.

Promoting freedom of speech: While hate speech is not protected by most laws, there is a need to balance the right to free speech with the need to protect individuals from harm [4]. Accurate hate speech detection can help identify harmful speech and prevent unnecessary restrictions on free expression.

Improving online communities: Hate speech has been shown to have a negative impact on online communities, driving away users and reducing engagement [5]. We can help create more welcoming and inclusive online environments by detecting and addressing hate speech.

Supporting law enforcement: Hate speech can be a precursor to hate crimes, which are illegal and can have serious consequences [6]. By detecting hate speech, law enforcement can better identify potential threats and take appropriate action to prevent violence.

In this paper, we discuss a shared task "HOMO-MEX: Hate speech detection in Online Messages directed towards the Mexican Spanish-speaking LGBTQ+ population". According to the task organizer, the LGBT+ community is disproportionately affected by mental health issues, substance addiction disorders among its members, discrimination in the labor markets, and being denied access to education and health services. Even though there have been significant strides made around the world to combat this form of discrimination, the vast majority of the LGBT+ population continues to struggle with the effects of LGBT+phobia. Given this scenario, they proposed this shared task, the objective of which is to enhance the performance of automatic detection systems developed for the purpose of classifying hate speech directed at the LGBT+ population [7].

This paper is structured into six sections. The first section provides an overview of the current state of the field under investigation. Moving on, the second section focused on describing the specific task that the paper aims to address. The third section delves into the dataset used in the research. The fourth section delves into the challenges associated with the task. Moving forward, the fifth section presents the experiment or methodology employed to address the task.Lastly, the paper concludes by summarizing the proposed solution for the given task.

2. Related Work

Nayak and Joshi [8] examined the use of transformer-based methods for analyzing code-mixed English-Hindi text, utilizing parent tweets as contextual information. The performance of multilingual BERT and Indic-BERT models were evaluated in both single-encoder and dual-encoder settings. In the single-encoder approach, the target and context texts were concatenated and fed into the BERT model, while the dual-encoder approach encoded the two texts independently and averaged the resulting representations. The study found that the dual-encoder approach with independent representations produced better results. The models were trained using the PyTorch framework and Hugging Face library, and fine-tuned for up to 5 epochs, achieving a maximum F1 score of 73.07% on the mixed dataset.

Arif et al. [9]explored various algorithms for detecting fake news in multiclass and cross-

lingual settings. To assess the effectiveness of these algorithms, the author reported macro F1 scores for both mono-lingual and cross-lingual tasks. For the mono-lingual task in English, the RoBERTa pre-trained model was employed, and a macro F1-score of 28.60% was achieved. In the cross-lingual task, the Bi-LSTM deep learning algorithm was utilized for detecting fake news in both English and German. The resulting macro F1-score was 17.21%. These scores suggest that detecting fake news across multiple languages presents significant challenges for NLP models, as they must contend with issues such as differing syntactical structures.

Tiţa and Zubiaga [10]presented an illustration of the capabilities of fine-tuned altered multi-lingual Transformer models (mBERT, XLM-RoBERTa) in regard to essential social data science tasks with cross-lingual training to detect hate speech from English to French and vice versa, and each language on its own. The paper also includes iterative improvement and comparative error analysis.

Mozafari et al. [11]found that BERT is particularly useful for monolingual multi-class hate speech classification on Twitter, showcasing superior performance compared to similar approaches. However, the authors also noted two significant challenges in this task, namely the limited availability of labeled data and the presence of bias. Despite these challenges, the study reported good performance metrics. This research emphasized the potential of BERT and its descendants and suggested further experimentation with its architecture and embeddings to improve its capabilities. Overall, the study highlighted the importance of using BERT in social media analysis and the need for ongoing research to address the challenges in this domain.

Tonja et al. [12]investigated the feasibility of using a language-specific pre-trained language model to identify aggressive and violent incidents in Spanish social media for the DAVINCIS: @IberLEF2022 shared task. The study employed a distilled version of BETO called DistilBETO and optimized the model using an Adam optimizer with a batch size of 64 and a learning rate of 0.0001. The maximum number of epochs was set to 10, and early stopping was based on the validation set's performance. The researchers also applied a dropout rate of 0.2 to regularize the model. The proposed model achieved an F1 score of 74.55% for violent events on the DA-VINCIS dataset.

3. Task Description

As stated in the introduction, this work aims to identify hate speech directed at the LGBT+ community. The task includes two tasks.

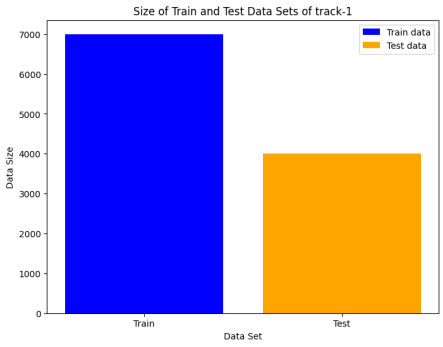
Track-1: This task involved identifying hate speech using a multi-class approach. The three classes used were meant to determine whether the content contained any LGBT+phobic material or not. These classes included LGBT+phobic (P), not LGBT+phobic (NP), or not related to LGBT+ (NA).

Track-2: This subtask is designed to detect fine-grained multi-label hate speech which indicates a kind of LGBT+phobia in the given tweets. Specifically, the algorithm is intended to detect different types of phobias: Lesbophobia (L), Gayphobia (G), Biphobia (B), Transphobia (T), and other forms of LGBT+phobias (O).

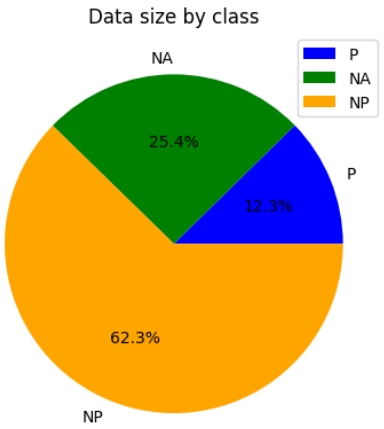
4. Data Collection and Preparation

The subtasks are reliant on the HOMO-MEX corpus, which served as the primary data source. The corpus comprises original tweets that were extracted from a period spanning from 2012 to 2022, and each tweet was manually annotated. In order to provide comprehensive information for each tweet in the corpus, we included details such as the classification of the tweet into one of three categories, namely P, NP, or NA. Additionally, we labeled each tweet according to the type of phobia it represents, which includes L, G, B, T, and O. In some datasets, it is possible for each data point to have more than one label associated with it. This means that a data point can belong to multiple labels simultaneously, rather than being assigned to a single, mutually exclusive label. This is commonly referred to as multi-label classification.

In Track-1, we were provided with a dataset containing 7000 data for training and 4000 samples for testing. Upon analyzing the training data, we discovered that it could be classified into three distinct categories: P, NA, and NP. The number of samples belonging to each category were found to be 862, 1778, and 4360 respectively. Figure 1 shows information provided that will be useful in developing and evaluating the proposed models for the task at hand.



(a) Train and test data size in track 1



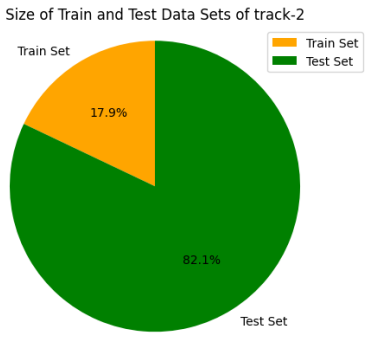
(b) Multi-class size in training data of track 1

Figure 1: Data statistics in track 1

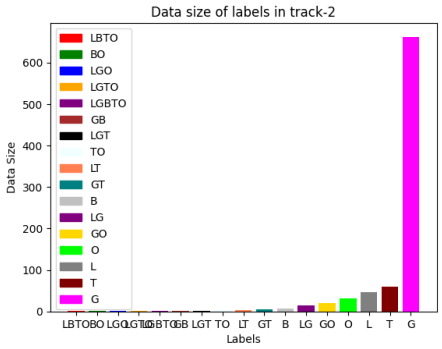
Track-2 includes training and testing data with 862 and 3941 datasets, respectively. On the other hand, we analyzed the training data which contains various fine-grained multi-labels. Figure 2 depicts the statistics of the given data.

4.1. Pre-processing

Pre-processing is a critical step in preparing the data for training and testing the models [13]. It can have a significant impact on the accuracy and generalization of the model, as well as the efficiency of the training process. The provided dataset was raw data, which means it



(a) Train and test data size in Track 2



(b) Multi-label size in training data of Track 2

Figure 2: Data statistics in Track 2

needs pre-processing to be clear and accurate. Cleaning unwanted data means we remove stopwords, HTML tags, URLs, digital numbers, and non-alphabetic characters and convert them to lowercase. Also, we defined a function to replace emoji characters with their meaning.

In order to address the data imbalance problem, we used an oversampling technique in Track 1, which is random oversampling that involves increasing the size of the minority class to match the size of the majority class. Therefore, random samples are duplicated in the minority class. In Track 2, we did not apply any techniques to balance the given data. Imbalance usually leads to overfitting and causes the model to perform poorly on test data [14]. There are large variations in the data for each label, for instance, as we can see in Figure 2b, the data size based on their label is XGXXX=662, XXXTX=59, XXXXO=31, XGXXO=21, LGXXX=15, XGXTX=5, and the rest of the labels are less than 5. Training data in Track 1 is already ordered in some way. In order to remove any existing order bias, we shuffled the rows.

5. Challenges of the task

As we discussed in section 4 the above section and observed from Figures 1 and 2, there are three problems: First, in both tracks, the training data was not balanced, which leads to bias in the model's performance. Next, in Track 1 the training data is sorted by label, which could lead to a model that is overfitting to certain labels and underfitting to others [14]. In order to resolve those problems, we used techniques discussed in section 4. In Track 2 the size of the test data is greater than the size of the training data. With a smaller training dataset, the model may not be able to learn the underlying patterns and relationships in the data, leading to poor performance on the test set [15].

6. Experiments and Results

We reviewed different related works in order to understand the state of the arts in this area. Various researchers used different techniques such as KNN [16], CNN [17], LSTM[18], BiLSTM [19], and transformer-based approaches [20].

For the tasks, we chose the transformer-based BERT and RoBERTa-based approaches. Both are state-of-the-art pre-trained language models for NLP tasks, developed by researchers at Google and Facebook respectively [21, 22]. In the experiments, we used RoBERTa transformer-based pre-trained model to detect multi-class hate speech Track 1 and BERT transformer-based pre-trained model to identify fine-grained multi-label hate speech.

Both architectureS of a pre-trained RoBERTa and BERT models were implemented for sequence classification with 3 classes and 18 labels respectively. The models consist of an embedding layer, an encoder layer, and a classification head. Figure 3 illustrates both transformer-based model parameters we used in the experiment.

```
name             module
-----------------------
roberta:embeddings
roberta:encoder
classifier       RobertaClassificationHead(
  (dense): Linear(in_features=768, out_features=768, bias=True)
  (dropout): Dropout(p=0.1, inplace=False)
  (out_proj): Linear(in_features=768, out_features=3, bias=True)
)
```

(a) Roberta transformer parameters

```
name             module
-----------------------
bert:embeddings
bert:encoder
bert:pooler
dropout          Dropout(p=0.1, inplace=False)
classifier       Linear(in_features=768, out_features=18, bias=True)
```

(b) Bert transformer parameters

Figure 3: Transformer-based model parameters

The embedding layer converts the tokenized input sequence into a fixed-size vector representation for each token. The encoder layer consists of a stack of transformer blocks that process the input sequence and capture its contextual information. Finally, the classification head takes the output of the encoder and maps it to its output labels using a linear layer with a softmax activation function. We used the tokenizer from the transformer library to tokenize the input data. The tokenized data was transformed into a tensor to get the token, segments and label tensors. Figure 4 depicts how the model classifies hate speech.

RoBERTa and BERT transformer-based pre-trained model were trained trains in 5 and 10 epochs, respectively using Track 1 and 2 training data. In both models, we used Adam optimizer with a learning rate of 1e-5. The models performed zero-padding on the tokens tensors and segments, tensors. Then they were tested on a separate dataset and evaluated with F1 score. In Track-1, the RoBERTa pre-trained model achieved an F1 score of 79.59%, while the BERT model performed better in Track-2, achieving an F1 score of 67.33%.

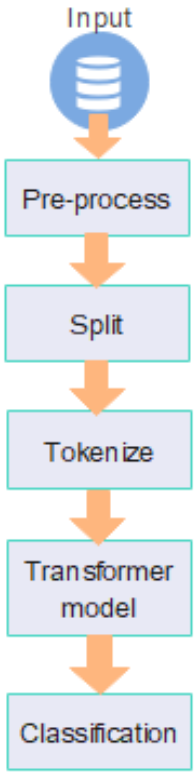


Figure 4: Experimental architecture for Transformer based hate speech detection

The findings showed that the RoBERTa transformer pre-trained model better effectively detected hate speech with multiple classes. On the other hand, the BERT transformer pre-trained model outperformed in detecting finer-grained multi-labels, as it only classified into 18 labels. Fine-tuned Roberta model for multi-class can be obtained at the Hugging Face website (https://huggingface.co/Mesay/Homo-mex-multi-class-hate-speech), or the fine-tuned BERT model for multi-label classification (https://huggingface.co/Mesay/Homo-mex-multi-label-hate-speech).

7. Conclusion

In this paper, we explored the feasibility of using a transformer-based pre-trained model to categorize hate speech into multiple classes and labels. In order to train the model, we first pre-processed the available data, which was a time-consuming operation due to the poor quality of the data. To identify hate speech, we performed two experiments. One experiment utilized the RoBERTa model and was more effective at identifying multi-class hate speech. We trained the model for 10 epochs and achieved an F1 score of 79.59%. The second experiment used the

BERT model to detect fine-grained, multi-label hate speech. This experiment achieved an F1 score of 67.33% with 5 epochs. We also discussed the difficulties we encountered during the experiments and resolved them by utilizing various techniques that improved our accuracy. We conclude that transformer-based approaches are suitable for identifying hate speech directed at the LGBT+ community.

We suggest that researchers who work in the task of hate speech detection pay close attention to data preparation and pre-processing including balancing the data, as it can significantly affect the model's performance. Furthermore, we recommend conducting more analyses and experiments using different techniques.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 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 Ph.D. Award.

References

[1] Z. Mossie, J.-H. Wang, Social network hate speech detection for amharic language, Computer Science & Information Technology (2018) 41–55.

[2] H. M. Saleem, K. P. Dillon, S. Benesch, D. Ruths, A web of hate: Tackling hateful speech in online social spaces, arXiv preprint arXiv:1709.10159 (2017).

[3] J. T. Nockleby, Hate speech in context: The case of verbal threats, Buff. L. Rev. 42 (1994) 653.

[4] J. Waldron, The harm in hate speech, Harvard University Press, 2012.

[5] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, M. Madden, Social media update 2014, Pew research center 19 (2015) 1–2.

[6] K. Crawford, M. L. Gray, K. Miltner, Big data| critiquing big data: Politics, ethics, epistemology| special section introduction, International Journal of Communication 8 (2014) 10.

[7] 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).

[8] R. Nayak, R. Joshi, Contextual hate speech detection in code mixed text using transformer based approaches, arXiv preprint arXiv:2110.09338 (2021).

[9] M. Arif, A. L. Tonja, I. Ameer, O. Kolesnikova, A. Gelbukh, G. Sidorov, A. G. M. Meque, Cic at checkthat! 2022: multi-class and cross-lingual fake news detection (2022).

[10] T. Tiţa, A. Zubiaga, Cross-lingual hate speech detection using transformer models, arXiv preprint arXiv:2111.00981 (2021).
[11] M. Mozafari, R. Farahbakhsh, N. Crespi, A bert-based transfer learning approach for hate speech detection in online social media, in: Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8, Springer, 2020, pp. 928–940.
[12] A. L. Tonja, M. Arif, O. Kolesnikova, A. Gelbukh, G. Sidorov, Detection of aggressive and violent incidents from social media in spanish using pre-trained language model, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS. org, 2022.
[13] 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, arXiv preprint arXiv:2211.14459 (2022).
[14] W. Koehrsen, Overfitting vs. underfitting: A complete example, Towards Data Science (2018) 1–12.
[15] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, et al., Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography 36 (2013) 27–46.
[16] M. S. 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, 2022, pp. 25–28.
[17] Q. Huang, R. Chen, X. Zheng, Z. Dong, Deep sentiment representation based on cnn and lstm, in: 2017 international conference on green informatics (ICGI), IEEE, 2017, pp. 30–33.
[18] M. G. Yigezu, M. M. Woldeyohannis, A. L. Tonja, Multilingual neural machine translation for low resourced languages: Ometo-english, in: 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), IEEE, 2021, pp. 89–94.
[19] M. G. Yigezu, A. L. Tonja, O. Kolesnikova, M. S. 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, 2022, pp. 29–33.
[20] A. L. Tonja, O. E. Ojo, M. A. Khan, A. G. M. Meque, O. Kolesnikova, G. Sidorov, A. Gelbukh, Cic nlp at smm4h 2022: a bert-based approach for classification of social media forum posts, in: Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task, 2022, pp. 58–61.
[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[22] 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).