Hate Speech Detection Against the Mexican Spanish LGBTQ+ Community Using BERT-based Transformers

Carlos Fernández Rosauro[1,*], Montse Cuadros[2]

[1]Universitat Oberta de Catalunya, Barcelona, Spain
[2]SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi Pasealekua 57, Donostia/San-Sebastián, 20009, Spain

Abstract

In this paper we present our approach to the HOMO-MEX task: Hate speech detection in Online Messages directed tOwards the MEXican spanish speaking LGBTQ+ population. We present our results for both Track 1: Hate speech detection track, in which the aim is to indicate whether a set of tweets exhibit LGBT+phobic content or not, and Track 2: Fine-grained hate speech detection track (Multi-labeled), in which the tweets labeled as LGBT+phobic need to be classified according to the type of LGBT+phobia they show. We utilized both classical machine learning and Transformer-based deep learning models focused on BERT-like architectures to tackle both tracks. The model that achieved the best results in terms of F1-Score (0.84 in Track 1) and macro-average F1-Score (0.68 in Track 2) was robertuito-base-uncased. With this model our team reached the 2nd position in both tracks.

Keywords

NLP, Text Classification, Hate Speech, Deep Learning

1. Introduction

This paper describes our participation in HOMO-MEX 2023 shared task, which is part of the iberLEF Conference. This challenge is focused on Hate speech detection in Online Messages directed tOwards the MEXican spanish speaking LGBTQ+ population. It is split in two tasks:

- In Track 1, participants have been provided with a collection of tweets and tasked with determining whether these tweets contain LGBT+phobic content. The classification options available for the tweets are: LGBT+phobic (P), not LGBT+phobic (NP), or not LGBT+related (NA).
- Track 2, on the other hand, involves fine-grained hate speech detection. Participants are required to identify the specific type of LGBT+phobia conveyed in the labeled tweets. The classification categories for this track include Lesbophobia (L), Gayphobia (G), Biphobia (B), Transphobia (T), and/or other forms of LGBT+phobia (O). Participants have the flexibility to assign one or more labels to each tweet. For instance, a tweet might be

IberLEF 2023, September 2023, Jaén, Spain
*Corresponding author.
cfernandezrosa@uoc.edu (C. F. Rosauro); mcuadros@vicomtech.org (M. Cuadros)
https://github.com/cfernandezros/ (C. F. Rosauro)
0000-0002-3620-1053 (M. Cuadros)

assigned multiple labels such as "L", "G", and "B" while another tweet might only receive the label "O".

We refer the reader to the overview article [1] of the HOMO-MEX 2023 competition for further information. Our team has participated in both tasks, implementing baseline models and more sophisticated systems based on BERT-like [2] Transformers. Both tasks are related to text classification tasks where mainly differ in that Track 1 is a multi-class problem and Track 2 is a multi-label problem.

This paper is organized as follows, section 2 presents the datasets available for the shared task which have been used. Section 3 presents the basic system used for both tracks. Section 4 presents the experimentation performed in Track 1 and in Track 2. Finally, section 5 draw some concluding remarks of our participation in the shared task.

2. Dataset

In the shared task, the official dataset for Track 1 regarding training data contains 7000 tweets labelled as either LGBT+phobic (P), not LGBT+phobic (NP) or not LGBT+related (NA) sorted by an index. Regarding Track 2, training dataset contains 862 LGBT+phobic tweets along with 5 columns that define the type of hate speech that the tweet contains in a binary-variable manner (1 or 0). This way, Track 2 tweets can contain one or multiple of the following behaviours: Lesbophobia (L), Gayphobia (G), Biphobia (B), Transphobia (T), and/or other LGBT+phobia (O).

Track 1 training set is split into 80% for training and 20% for validation, while Track 2 training set is split into 85% for training and 15% for validation . In both cases this was done by randomly selecting the validation split using a preset seed. Table 1 shows the number of tweets per track in train and validation sets and the average word tokens per tweet.

Table 1
Training dataset for Track 1 and Track 2 in number of tweets and avg. word tokens per tweet

	Track 1	Track 2
Train	5600	1400
Validation	732	132
Avg. word tokens	21.7	16.2

3. Methodology

For both tasks, the same methodology was followed: first, two classical methods were implemented as a baseline and then, a series of BERT-based models were utilized to search for better results.

In the next subsection we present the Baseline models and the BERT-based models.

3.1. Baseline Models

Baseline models are both learnt on a TF-IDF matrix generated through the TfidfVectorizer class from the sklearn library. On each task, the vectorizer is fitted on the training dataset and then transformed on both the training and validation datasets. This training and validation TF-IDF matrixes are then used to train and validate the baseline methods. Before the TF-IDF matrixes are generated, the tweets follow a processing pipeline, which includes a Snowball Stemmer from the NLTK library.

- **Multinomial Naive Bayes**[3]**.** From the sklearn Python library, fitted on the input TF-IDF matrix and output labels with default parameters.
- **Linear SVC**[4]**.** From the sklearn Python library, fitted on the input TF-IDF matrix and output labels with a linear kernel and all other default parameters.

3.2. BERT-based Models

Transformer models are implemented by downloading the pre-trained base models from Hugging Face and then fine-tunning them on the text data using Tensorflow, Keras and the Transformers library.

The text data is not processed as before in the case of these models but with their own built-in tokenizers through the Transformers library. Additionally, in the case of RoBERTuito, the tweets need to be pre-processed with the pysentimiento library.

- **bert-base-spanish-wwm-cased (BETO).** BETO[5] is a Spanish version of the BERT-Base model.
- **robertuito-base-uncased.** RoBERTuito[6] is a RoBERTa implementation for social media text in Spanish trained on 500 million tweets.
- **mdeberta-v3-base.** mDeBERTaV3[7] is a multilingual implementation of the DeBERTa architecture.

4. Experimentation

4.1. Track 1

In this task the goal is to train and validate each of the selected models to a multi-class problem where the input data is the text data (tweets) from Track 1 and the output data are the assigned labels.

As explained in section 3, we have participated in this task with two baselines and three BERT-based Transformer models.

When working with the BERT-based models, the checkpoints are downloaded from their repository at Hugging Face and then loaded as Tokenizers that are used to process both training and validation tweets. The model is then instantiated using the TFAutoModelForSequenceClassification class from the Transformers library using the same checkpoint and the parameters and hyperparameters are tuned in order to get the best results. The TFAutoModelForSequenceClassification class is adapted to a multi-class problem. We have tuned the number of epochs,

batch size, start and end learning rates on the polynomial scheduler and the dropout probability as shown in Table 2.

Then the model is compiled and fitted with Tensorflow and Keras. In the compilation step, the AdamW[8] algorithm is always applied with a polynomial learning rate scheduler. After the model is fitted to the training data, it is validated to the validation dataset using weighted-average F1-Score, Precision and Recall and Accuracy.

Regarding performance, it is worth mentioning that all Transformer models are fine-tuned in Google Colab using NVIDIA Tesla T4 or V100 GPUs expect for the mDeBERTaV3 model, for which an NVIDIA A100 is utilized.

In the case of RoBERTuito, as we figured that this model displayed more tolerance than the others to tuning techniques, we added a class weight dictionary {NP: 0.7, NR: 1, P: 1.3} for the training phase to take care of the class imbalance situation that specially affects the model's ability to identify LGBT+phobic (P) tweets (minority class). This set of weights was manually tweaked.

Table 2
Experimentation set-up for Track 1

Model	Epochs	Batch size	start Learning rate	end Learning rate	dropout
BETO	4	4	$5e^{-5}$	0	-
RoBERTuito	7	4	$2e^{-5}$	0	0.2
mDeBERTaV3	5	32	$5e^{-5}$	0	0.2

As shown in table 3 weighted-average F1-Score, Precision and Recall and Accuracy were used to benchmark the models implemented in Track 1. RoBERTuito achieved the best results across all metrics, though neither this model nor any other Transformer-based model achieved a significant improvement over the simple Linear SVC baseline trained on TF-IDF matrixes, which doesn't take into account word order or semantic similarity between tokens as Transformer-based models do.

Table 3
Track 1 results obtained on validation data based on weighted F1-score, Precision and Recall and Accuracy

Model	w.F1-Score	w.Precision	w.Recall	Accuracy
Multinomial Naive Bayes	0.76	0.94	0.68	0.68
Linear SVC	0.84	0.85	0.83	0.83
BETO	0.86	0.87	0.86	0.86
RoBERTuito	**0.88**	**0.88**	**0.88**	**0.88**
mDeBERTaV3	0.83	0.83	0.84	0.84

4.2. Track 2

In this track the goal is to train and validate each of the selected models to a multi-label problem where the input data is the text data (tweets) from Track 2 and the output data are the multiple assigned labels.

Similar to Track 1, we have worked using the same procedure using Hugging Face models and the Tokenizer class. In the same way, we use TFAutoModelForSequenceClassification class using the parameters and hyperparameters showed in table 4 and adapt them to the multi-label problem. Then the model is compiled and fitted with Tensorflow and Keras. In the compilation step, the AdamW[8] algorithm is also applied with a polynomial learning rate scheduler. Regarding this track, when the model is fitted to the training data, it is validated to the validation dataset using macro-average F1-Score, Precision and Recall and Accuracy.

As for Track 1, we have participated in this task with two baselines and three BERT-based Transformer models. In the case of performance, we have also fine-tuned all Transformer models in Google Colab with the same specifications followed in Track 1.

Table 4
Experimentation set-up for Track 2

Model	Epochs	Batch size	start Learning rate	end Learning rate	dropout
BETO	4	4	$5e^{-5}$	0	0.2
RoBERTuito	7	4	$2e^{-5}$	0	-
mDeBERTaV3	5	32	$5e^{-5}$	0	-

As shown in table 5, macro-average F1-Score, Precision and Recall and Accuracy were utilized to benchmark the models implemented in Track 2. RoBERTuito achieved the best result on Accuracy and BETO achieved the best results on macro-average Precision and Recall. Again, none of the Transformer-based models achieved a significant improvement over the simple Linear SVC baseline.

Table 5
Track 2 results obtained on validation data based on macro-average F1-score, Precision and Recall and Accuracy

Model	macro F1-Score	macro Precision	macro Recall	Accuracy
Multinomial Naive Bayes	0.18	0.20	0.17	0.78
Linear SVC	0.49	0.44	0.59	0.87
BETO	0.52	0.51	0.59	0.86
RoBERTuito	**0.52**	**0.48**	**0.58**	**0.88**
mDeBERTaV3	0.19	0.20	0.19	0.78

4.3. Discussion on Results

We have seen how in both tasks a simple Linear SVC model almost achieved the same results as our best performer model RoBERTuito. We didn't submit Linear SVC to the HOMO-MEX competition so we can only conclude inside our experimental setup that simpler models such as a Linear SVC will perform as good as Transformer models if properly tuned because in the context of short social media texts, in this experimentation, language models don't seem to outperform in the classification problem. We can conclude then, that the existence (or lack of) LGBT+phobic terms and their type are the most important features in the context of both classification tasks, as the performance of such a simple statistical method over term frequencies demonstrates.

5. Conclusions

We have presented our participation in the shared task Homo-Mex 2023. We have participated in both tracks obtaining a second position according to official metrics reported by organizers. We have presented a set of experiments using classic algorithms and BERT-like Transformer models. The results obtained in both tasks exhibit similar performances for different methodologies.

6. Acknowledgements

This work was undertaken by the principal author in fulfilment of his Bachelor's Final Project at the Universitat Oberta de Catalunya.

References

[1] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S.-T. Andersen, S.-L. Ojeda-Trueba, Overview of HOMO-MEX at Iberlef 2023: Paraphrase Detection in Spanish Shared Task, Procesamiento del Lenguaje Natural 71 (2023).

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[3] C. Manning, P. Raghavan, H. Schuetze, Introduction to information retrieval, Cambridge University Press, 2008, pp. 234–265. URL: https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html.

[4] C. Manning, P. Raghavan, H. Schuetze, Introduction to information retrieval, Cambridge University Press, 2008, pp. 319–325. URL: https://nlp.stanford.edu/IR-book/html/htmledition/support-vector-machines-the-linearly-separable-case-1.html.

[5] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[6] 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.
[7] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. arXiv:2111.09543.
[8] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.