Mixture Models based on BERT for Hate Speech Detection

Haoyang Chen, Zhongyuan Han*, Leilei Kong, Zhijie Zhang, Zengyao Li, Mingcan Guo, Haoliang Qi

Foshan University, Foshan, China

Abstract

While social platforms such as Twitter have brought convenience to people, they have also become a hotbed for spreading hate speech. Identifying hate speech and offensive content has become an important task. This paper presents our team's experiments on two shared tasks of HASOC 2022, where we fine-tuned three pre-trained models based on indic-abusive and multilingual BERT to perform hate speech detection on tweets in code-mixed languages. We try to reduce the impact of data imbalance by combining model predictions. Our team obtained 5th (with macro f1: 0.6388) in the dichotomous subtask 1 for Hinglish and German and 3rd (with macro f1: 0.4769) in subtask 2 for Hinglish with multiple classifications.

Keywords

Hate Speech Detection, BERT, Classification, Hinglish

1. Introduction

With the widespread popularity of social media platforms such as Twitter and Facebook worldwide, users are free to express their thoughts and opinions. However, it has become a new challenge to detect and deal with the hate sentiment: the voice of hate speech and offensive content will cause severe mental stress to the victims, lead to social tensions, and lead to confrontation and violence [1, 2]. Similar objectionable content has seriously affected people's daily lives, and there is an urgent need to find a low-cost way to solve this challenge.

As a result, social media companies such as Twitter and YouTube have developed their detection systems to monitor user posts and filter hate content. However, current detection systems are mainly targeted at English-speaking environments and are still less practical for languages excluding English or Code-Mixed languages, such as Hinglish. In addition, it remains a chronic problem where contextual information is needed to identify hate speech (e.g., comments that do not contain hate per se but identify with parent tweets that are hate speech). In this state, Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) proposes a series of hate detection classification tasks for low-resource languages, aiming to improve hate detection under low-resource languages. This paper describes our team's (fosu-nlp) working notes for the HASOC 2022 subtask. We studied several pre-trained BERT models and tried to combine them to accomplish the task.

The rest of the paper is organized as follows: Section 2 provides an overview of recent works in hate detection. Section 3 briefly describes the task and dataset composition of HASOC 2022. We will present our methodology and model in Section 4, and the model results will be presented in Section 5. Finally, Section 6 summarizes our work.

2. Related Works

Many effective methods have been proposed for hate speech detection in recent years. Gambäck et al. [3]proposed to classify hate speech using CNN for word2vec embedded Twitter texts. Ayo et al. [4]

Forum for Information Retrieval Evaluation, December 9-13, 2022, India

EMAIL: hoyo.chen.i@gmail.com (H. Chen); hanzhongyuan@gmail.com (Z. Han)(*corresponding author); kongleilei@fosu.edu.cn (L. Kong); zhangzhijie5454@gmail.com (Z. Zhang); lzy1512192979@gmail.com (Z. Li) gmc9812@163.com (M. Guo); haoliang.qi@gmail.com (H. Qi)
ORCID: 0000-0003-3223-9086 (H. Chen); 0000-0001-8960-9872 (Z. Han); 0002-4636-3507(L. Kong); 0000-0002-4854-0618 (Z. Zhang); 0000-0001-8472-4150 (Z. Li); 0000-0002-4977-2138 (M. Guo); 0000-0003-1321-5820 (H. Qi)

performed feature extraction and topic inference of Twitter tweets by TF-IDF and Bayes classifier and proposed a rule-based clustering model. Furthermore, various attempts based on SVM [5], LSTM [6], or the current state-of-the-art pre-trained BERT [7] have been proposed for hate detection in mixed English and Hindi languages.

The HASOC committee has organized a series of tasks in the last few years [8, 9, 10]. At HASOC 2021, Banerjee et al. [11] explored some fine-tuning Transformer models and designing a weighted classifier layer at the final phase. Bhatia et al. [12] utilized an Emoji2Vec system to convert emojis into vectors to add features using emoji data instead of simply removing them. Regarding context-based hate speech detection, Zaki et al. [13] generated results by obtaining predictions from three BERT models and using soft/hard voting, and they ended up with an f1 score of 0.7253. It can be seen that pre-training-based models have significant potential, so we will continue to investigate the application of pre-training models for hate speech detection.

3. HASOC Task and Datasets

At this year's Forum for Information Retrieval Evaluation (FIRE, 2022), HASOC brought a new set of shared tasks [14], including identifying hate-speech posts in a code-mixed language on Twitter. Our team focused on subtasks 1 and 2: Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) [15] Binary and Multiclass. Subtask 1 is a coarse-grained binary classification task that aims to identify hate speech in German and Hinglish (Hindi and English) conversations. Tweets should be classified with the following tags:

- (NOT) Non- Hate-Offensive: The tweet does not contain hate speech.
- (HOF) Hate and Offensive: The tweet reflects hateful, offensive, or profane content.

As an extension of subtask 1, subtask 2 is a triple classification task that further classifies the tweets as follows:

- (SHOF) Standalone Hate: The tweet, comment, or reply contains hateful, offensive, and profane content.
- (CHOF) Contextual Hate: The comment or reply is treated as hate speech by supporting the hateful content expressed in its parent. This includes affirming the hate speech with a positive sentiment.
- (NONE) Non-Hate: The tweet, comment, or response is not hate speech.

The dataset given by HASOC was sourced from Twitter and provided data on tweets and their replies in German and a code-mixed of English and Hindi languages. Statistical information on the number of labels in the original dataset is provided in Table 1.

Table 1
Statistics of the original dataset

Class	Hinglish	German
NOT	2390	219
HOF	2524	88
SHOF	1636	-
CHOF	888	-
NONE	2390	-
Total	4914	307

The dataset was stored as a tree structure. For the model to obtain contextual information, the conversation must be flattened and stitched as the "parent-comment-reply" chain. Considering the actual situation that the model does not know the content of the parent tweets at the time of prediction,

we choose to divide the original tweet nodes in a 9:1 ratio to form a new training and validation set. These datasets will be flattened in the next step.

4. Methodology

In this paper, two BERT models were used for fine-tuning experiments. The one is indic-abusive-allInOne-MuRIL [16], a newly proposed hate-detection-binary-classification-model trained for Indian multilingual by team Hate alert, which will be utilized as the primary model for the Hinglish binary task. The other is multilingual BERT [17], which will be used to handle the German part of the binary classification as well as the task of multiclassification.

4.1. Data Pre-processing

First, we expand the tweets and link them to the corresponding tags. Considering the maximum input length limit of BERT, the conversation set will be flattened in the reverse order as "reply-comment-parent," and all data will be pre-processed, specifically:

- All @USER and URL will be removed
- Extra spaces and line feeds will be removed
- All tweets will be normalized by stemming
- Stop word list will be applied to all the tweets

4.2. Subtask 1: ICHCL Binary

In the binary classification task of Subtask 1, two models were trained using tweet data from Hinglish and German, respectively. The Hinglish binary classification model (HNG-BCM), based on the fine-tuned indic-abusive pre-training model, aims to distinguish hate speech in Hindi and English mixed languages. The German part uses the fine-tuned multilingual bert as the German binary classification model (GER-BCM). The final output of both models will be combined and output as the result of Subtask 1.

4.3. Subtask 2: ICHCL Multiclass

For Subtask 2, we treated the triple classification as two associated binary classification tasks to reduce the effect of data imbalance. Two types of hate speech data in multi-label classification will first be used to train the standalone-contextual hate binary classification model (SCH-BCM), which to be able to distinguish between two different types of hate speech. The HNG-BCM in Subtask 1 will then perform the first classification on the test dataset. Then the perceived hate speech in the test dataset is sent to the SCH-BCM for a second classification to determine if it is contextual or standalone hate. In summary, HNG-BCM was used to determine if the input data was hate speech, and SCH-BCM focused on further differentiation of hate speech.

For the two subtasks, Figure 1 gives the corresponding flowcharts for each.

4.4. Experimental setting

For the two subtasks of HASOC 2022, our experiments used Hugging Face's transformer [18] library to fine-tune all pre-trained models. Those models were mostly configured with the same hyperparameters. The batch size was set to 32, and the maximum sequence length was 512. AdamW optimizer [19] with a linear learning rate scheduler and an initial learning rate of 2e-5 is used for training. For HNG-BCM and GER-BCM, we trained for 20 epochs, while for SCH-BCM, it is 40. Models will

be evaluated using macro f1 after training in each epoch. At the end of the training, the model with the highest score will be retained and used as the final model.

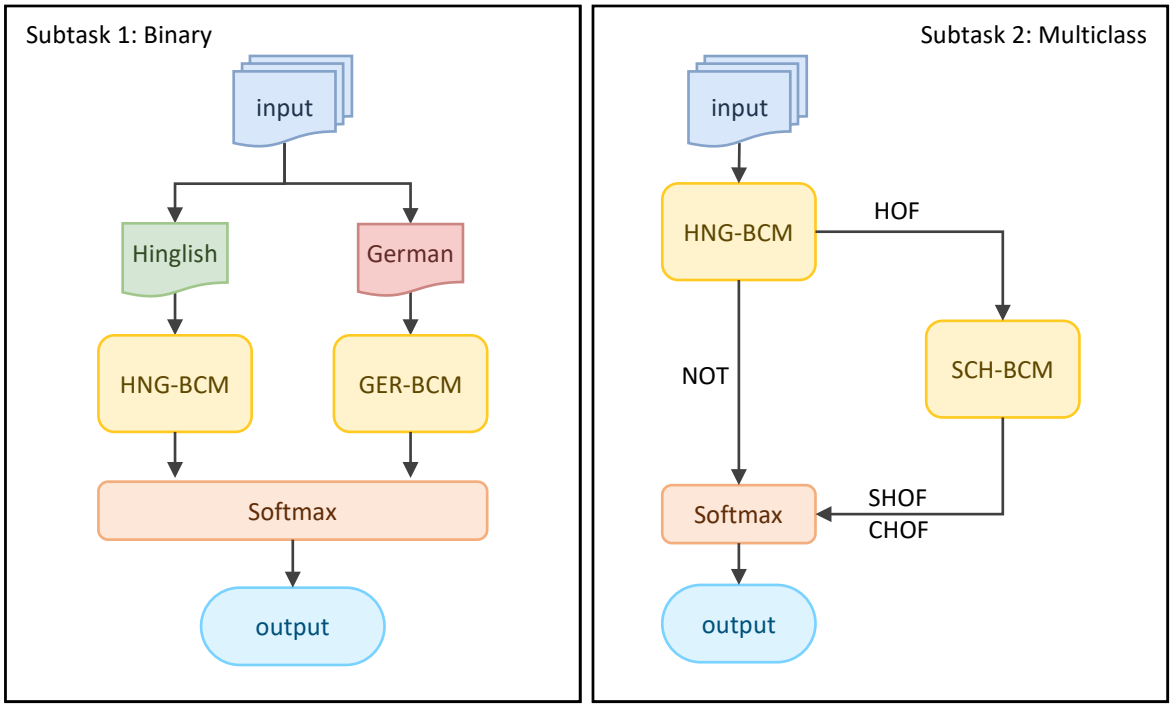

Figure 1: Task Flow Chart. In binary tasks, HNG-BCM (Hinglish binary classification model) and GER-BCM (German binary classification model) are fine-tuned for Hinglish and German, respectively. In multiclass tasks, SCH-BCM (standalone-contextual hate binary classification model) is designed to further distinguish between types of hate speech.

5. Results

We evaluate the model using the validation set, and macro f1 will be used as the evaluation metric. Table 2 shows the scores achieved by each model on the validation set.

Table 2
Macro F1 on the validation set

Model	Macro F1
HNG-BCM (Hinglish Binary)	0.6456
GER-BCM (German Binary)	0.6563
SCH-BCM (Hinglish Multiclass)	0.5661

In HASOC 2022, we ran three commits in total. The organizers used macro f1 to evaluate the predictions for each subtask. The final scoring metrics obtained by our model in the official test set can be found in Table 3.

Table 3
Metric on the test set

Task	Macro F1	Precision	Recall	Rank	1st Ranked Team / Macro F1
Task 1 ICHCL Binary	0.6388	0.6388	0.6388	5	nlplab_isi / 0.7083
Task 2 ICHCL Multiclass	0.4769	0.5042	0.4803	3	ub-cs / 0.4939

There are some differences between the final scores and the results on the validation set. This may be caused by the small validation set or the unevenness in partitioning the data, resulting in a higher score for the model on the validation set.

6. Conclusion

This paper briefly describes the results of our team's work on the HASOC 2022 shared task. Multiple pre-trained models have been used and processed in combination to solve the problem of hate speech detection based on the context of multilingual mixed tweets, and our team achieved competitive results for two subtasks. We note that the models perform poorly on the multiclassification task, likely to remain due to data imbalance. Our next work direction will consider trying to eliminate the data imbalance problem by adding training samples using multiple translations.

7. Acknowledgements

This work is supported by the Natural Science Foundation of Guangdong Province, China (No. 2022A1515011544).

8. References

[1] S. Jaki, T. De Smedt, M. Gwóźdź, R. Panchal, A. Rossa, G. De Pauw, Online hatred of women in the incels. me forum: Linguistic analysis and automatic detection, Journal of Language Aggression and Conflict, 2019, vol. 7, pp. 240–268.

[2] 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, pp. 88–93, 2016.

[3] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.

[4] F. E. Ayo, O. Folorunso, F. T. Ibharalu, I. A. Osinuga, A. Abayomi-Alli, A probabilistic clustering model for hate speech classification in twitter, Expert Systems with Applications, 2021, vol. 173, pp. 114762.

[5] K. Sreelakshmi, B. Premjith, K. Soman, Detection of hate speech text in hindi-english code-mixed data, Procedia Computer Science, 2020, vol. 171, pp. 737–744.

[6] T. Santosh, K. Aravind, Hate speech detection in hindi-english code-mixed social media text, in: Proceedings of the ACM India joint international conference on data science and management of data, 2019, pp. 310–313.

[7] A. Sharma, A. Kabra, M. Jain, Ceasing hate with moh: Hate speech detection in hindi–english code-switched language, Information Processing & Management, 2022, vol. 59, pp. 102760.

[8] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019, pp. 14–17.

[9] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Forum for Information Retrieval Evaluation, 2020, pp. 29–32.

[10] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech, in: Forum for Information Retrieval Evaluation, 2021, pp. 1–3.

[11] S. Banerjee, M. Sarkar, N. Agrawal, P. Saha, M. Das, Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages, in: FIRE (Working Notes), CEUR Workshop Proceedings, CEUR-WS.org, 2021, pp. 32–43.

[12] M. Bhatia, T. S. Bhotia, A. Agarwal, P. Ramesh, S. Gupta, K. Shridhar, F. Laumann, A. Dash, One to rule them all: Towards joint indic language hate speech detection, in: FIRE (Working Notes), CEUR Workshop Proceedings, CEUR-WS.org, 2021, pp. 419–431.

[13] Z. M. Farooqi, S. Ghosh, R. R. Shah, Leveraging transformers for hate speech detection in conversational code-mixed tweets, in: FIRE (Working Notes), CEUR Workshop Proceedings, CEUR-WS.org, 2021, pp. 63–74.
[14] S. Satapara, P. Majumder, T. Mandl, S. Modha, H. Madhu, T. Ranasinghe, M. Zampieri, K. North, D. Premasiri, Overview of the HASOC Subtrack at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: FIRE 2022: Forum for Information Retrieval Evaluation, Virtual Event, 9th-13th December 2022, ACM, 2022.
[15] S. Modha, T. Mandl, P. Majumder, S. Satapara, T. Patel, H. Madhu, Overview of the HASOC Subtrack at FIRE 2022: Identification of Conversational Hate-Speech in Hindi-English Code-Mixed and German Language , in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR, 2022.
[16] M. Das, S. Banerjee, A. Mukherjee, Data bootstrapping approaches to improve low resource abusive language detection for indic languages, in: HT, ACM, 2022, pp. 32–42.
[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, Association for Computational Linguistics, 2019, pp. 4171–4186.
[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, CoRR abs/1910.03771, 2019. URL: http://arxiv.org/abs/1910.03771. arXiv:1910.03771.
[19] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: ICLR, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.