AI_ML_NIT_Patna @HASOC 2020: BERT Models for Hate Speech Identification in Indo-European Languages

Kirti Kumari[a], Jyoti Prakash Singh[b]

[a]*Institute of Technical Education and Research (ITER), Siksha 'O' Anusandhan, Bhubaneswar, Odisha.*
[b]*National Institute of Technology Patna, Bihar, India.*

Abstract

The current paper describes the system submitted by team AI_ML_NIT_Patna. The task aims to identify offensive language in code-mixed dataset of comments in Indo-European languages offered for English, German, Hindi collected from Twitter. We participated in both Sub-task A, which aims to classify comments into two class, namely: Hate and Offensive (HOF), and Non- Hate and offensive (NOT), and Sub-task B, which aims to identify discrimination between Hate (HATE), profane (PRFN) and offensive (OFFN) comments. In order to address these tasks, we utilized pre-trained multi-lingual transformer (BERT) based neural network models and their fine-tuning. This resulted in a better performance on the validation, and test set. Our model achieved 0.88 weighted F1-score for English language in Sub-task A on testing dataset, and got 3rd rank on the leaderboard private test data having F1 Macro average of 0.5078.

1. Introduction

With the increase in usage of Internet, appealing to the propagation of thoughts, and expressions of an individual, there has been a tremendous increase in the spread of online hate speech [1, 2], cyberbullying [3], and cyber-aggression [4, 5]. Such phenomena have significantly affected the daily life of people, and such motives may also result in depression or suicide. It has become very important to track these behaviour in order to indicate activity that is hateful, offensive or that promotes violence towards an individual or group based on anything such as race, religion, gender or sexual orientation, to ensure that the Internet remains an open place, and to foster diversity of information, opinion, and innovation. It is very challenging task due to several reasons such as multi-linguality, multi-modality, non-standard writing style of social media post.

The Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) shared tasks of 2019, and 2020 focused on Indo-European languages in three different languages: English, German, and Hindi. The shared tasks have two sub-tasks: Sub-task A, and

Sub-task B focused on multi-lingual posts. Therefore, it gave us an opportunity to address the multi-lingual issues associated with social media posts. In order to address this, we applied pre-trained transformer based neural network (BERT) models, which are publicly available in multiple languages, and the model supports fine-tuning for specific tasks. In addition to this, its multi-lingual feature allows for us to analyse sentiment for the comments which have multiple language words, and sentences, for example HINGLISH (which is combination of English, and Hindi, commonly used as an expression medium) in Indian social media. We participated for both sub-tasks, and all the three languages. We have tried several neural networks architecture but found that our BERT model is performing better than other models. We achieved 3rd rank for English Sub-task A.

The organization of the paper as follows: the previous works in these scenarios are discussed in Section 2, the data description is explained in Section 3. The models description and their training are presented in Section 4 and Section 5, respectively. The results of various experiments are explained in Section 6 and the current paper summarizes the most important findings in Section 7.

2. Related Works

Lot of researchers and practitioners from industry and academia have been attracted towards the problem of automatic identification of hate speech, cyberbullying, cyber-aggression and offensive languages.

The dynamics of the definition of cyberbullying, cyber-aggression, hate speech and its undoubted potential for social impact are addressed by authors [1, 2, 3, 6], particularly in online forums and social networking sites. Mishra and Mishra [7] developed BERT model for identification of hate, offensive and profane comments of multi-lingual English, Hindi and German languages from Facebook and Twitter social media platform. Mujadia et al. [8] used hybrid of different machine learning, and neural network based models for hate, offensive, and profane identification on multi-lingual comments of English, Hindi, and German languages collected from Facebook, and Twitter. They found that word, and character TF-IDF features with ensemble model of SVM, Random Forest, and Adaboost classifiers with majority voting is performing better than the other models. Wang et al. [9] used LSTM with attention model to identify hate, offensive and profane comments of English comments of Facebook and Twitter and found that the k-fold ensemble method is performing better than other methods. Kumari and Singh [10] have introduced a four-layer CNN model with three embedding techniques for detecting hate speech, and offensive content for multi-lingual HASOC 2019 text comments collected from Facebook, and Twitter.

3. Dataset

The data supplied by the organizing team, are the posts collected from Twitter. The posts are in the following three languages: English, German, and Hindi. The competition has two sub-tasks for each language, i.e. Sub-task A, and Sub-task B. Sub-task A consists of labelling with Hate and Offensive (HOF) if the content of the post contains any hate speech, otherwise the

Table 1
Dataset description

Dataset	Sub-task	Class	#training samples
English	A	NOT	1510
		HOF	1969
	B	NONE	1954
		PRFN	1375
		OFFN	632
		HATE	312
German	A	NOT	1466
		HOF	953
	B	NONE	1860
		PRFN	364
		OFFN	248
		HATE	266
Hindi	A	NOT	2116
		HOF	847
	B	NONE	2829
		PRFN	366
		OFFN	662
		HATE	424

label should be Not Hate-Offensive (NOT). Sub-task B has a more fine-grained data which also pointed to discriminate between Hate, Profane, and Offensive posts by labelling with HATE, PRFN, and OFFN. We utilized the testing dataset released by the organizers. The details of the collection, and labelling of datasets are discussed in paper [11, 12]. We presented the description of training dataset in Table 1 where number of samples of specific class are given.

4. Models

We used BERT as our pre-trained language model because of its success in recent as well as availability in multiple languages. We finely tuned the already trained model on our training dataset, and generated the submission model.

We used BERT, specifically BERT BASE UNCASED, which is a transformers model pre-trained on a large corpus of data in a supervised fashion. It has 756 encoding length with 12 transformer layers, 12 attention heads, and 110 million parameters. Sentences are tokenized by BertTokenizerFast. Before padding up to 256 sequence length, Tokens are converted into Sequences. Two tokens were added to each input text (CLS, and SEP) marking the beginning, and the end of the sequence. These sequences are passed into the model. After that, we used BERT BASE UNCASED model from TFBertModel. It contains vocabulary, and pre-trained model. We used optimizer as Adam with Categorical Cross-entropy for Sub-task B, and Binary Cross-entropy for Sub-task A as a loss function.

Table 2
Results on testing samples

Dataset	Sub-task	Weighted F1-Score	F1 Macro Average
English	A	0.88	0.88
	B	0.80	0.58
German	A	0.82	0.76
	B	0.74	0.45
Hindi	A	0.61	0.49
	B	0.63	0.27

Table 3
Results on private leaderboard samples

Dataset	Sub-task	F1 Macro Average
English	A	0.5078
	B	0.2298
German	A	0.4768
	B	0.2210
Hindi	A	0.4561
	B	0.2399

5. Training

We have trained the BERT model on English, German, and Hindi datasets of HASOC 2020 for Sub-task A. In addition to HASOC 2020 datasets, we also used HASOC 2019 datasets to train our models for Sub-task B. Training involves batch of size 64 with 20 epochs (on an average). Adam is used with the learning rate of 5e-05. For training, we have utilized 80% samples from training data provided by the organizing team, and rest 20% was used for validation. After training the model, we evaluated our model with the provided testing data. We experimented with different number of epochs, and analysed the training model with each epoch. It was observed that on an average after 20 epochs, the training accuracy did not improve rather it was decreasing, and hence the accuracy started to oscillate with its highest being at 20 epochs. The final leaderboard is calculated with approximately 15% of the private test data by the organizing teams. F1 Macro average was used which was for the deciding parameter that can be observed from the classification report.

6. Results

In this competition, teams are ranked on performance in terms of F1 Macro average on final leaderboard results. The Table 2 presents our results on testing samples provided by the organizing team in terms of Weighted F1-Score and F1 Macro average as performance metrics. The Table 3 presents the final leaderboard results for all the languages with their respective

sub-tasks. According to the official results in Sub-task A, the best Marco F1 score of our model are 0.5078, 0.4768, and 0.4561 respectively, for English, German, and Hindi languages ranked 3rd place for English Sun-task A.

7. Conclusion

In this paper, we presented our approach based on fine-tuning monolingual, and multi-lingual transformer networks (BERT model) to classify Twitter posts in three Indo-European languages: English, German, and Hindi, for hate-speech, and offensive content identification. Using transfer learning with the pre-trained BERT model on English Sub-task A, we achieved rank 3 on leaderboard private test data.

References

[1] 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, 2017, pp. 1–10.

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

[3] K. Kumari, J. P. Singh, Y. K. Dwivedi, N. P. Rana, Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach, Soft Computing 24 (2020) 11059–11070.

[4] K. Kumari, J. P. Singh, Y. K. Dwivedi, N. P. Rana, Aggressive Social Media Post Detection System Containing Symbolic Images, in: Conference on e-Business, e-Services and e-Society, Springer, 2019, pp. 415–424.

[5] K. Kumari, J. P. Singh, AI_ML_NIT_Patna@ TRAC-2: Deep Learning Approach for Multi-lingual Aggression Identification, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 113–119.

[6] K. Kumari, J. P. Singh, Identification of cyberbullying on multi-modal social media posts using genetic algorithm, Transactions on Emerging Telecommunications Technologies (2020) e3907. doi:10.1002/ett.3907.

[7] S. Mishra, S. Mishra, 3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for Hate Speech Identification in Indo-European Languages., in: FIRE (Working Notes), 2019, pp. 208–213.

[8] V. Mujadia, P. Mishra, D. M. Sharma, IIIT-Hyderabad at HASOC 2019: Hate Speech Detection., in: FIRE (Working Notes), 2019, pp. 271–278.

[9] B. Wang, Y. Ding, S. Liu, X. Zhou, YNU_Wb at HASOC 2019: Ordered Neurons LSTM with Attention for Identifying Hate Speech and Offensive Language., in: FIRE (Working Notes), 2019, pp. 191–198.

[10] K. Kumari, J. P. Singh, AI_ML_NIT Patna at HASOC 2019: Deep Learning Approach for Identification of Abusive Content, in: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (FIRE 2019, December 2019), 2019, pp. 328–335.

[11] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer,

Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages), in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, CEUR, 2020.

[12] 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.