NITP-AI-NLP@HASOC-FIRE2020: Fine Tuned BERT for the Hate Speech and Offensive Content Identification from Social Media

Abhinav Kumar[a], Sunil Saumya[b] and Jyoti Prakash Singh[a]

[a]National Institute of Technology Patna,India
[b]Indian Institute of Information Technology Dharwad, Karnataka, India
[a]National Institute of Technology Patna, Patna, India

Abstract

The current paper identifies the offensive and hate content in three datasets of English, Hindi, and German. The dataset appeared in *HASOC-2020 track*. A fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model is proposed to identify hate and offensive contents from the social media posts. The experimental results show that the proposed BERT model achieved significant performance in identifying hate and offensive content. For English subtasks A and B F1-score reported were 0.5031 and 0.1623, for Hindi subtasks A and B F1-score reported were 0.5300 and 0.0940 and for German subtasks A and B F1-score reported were 0.5109 and 0.1214 respectively.

Keywords

Hate speech, Offensive contents, Social media, BERT

1. Introduction

Modern computers and technology have increased the social media user base exponentially. Consequently, the growth in the number of users posts per day has also increased exponentially. The internet has become an integral part of our everyday lives in recent years [1, 2, 3]. Most of us are increasingly becoming a social network addict in this digital era and have spent time surfing through the various media all day long. According to the latest survey, internet users now spend over 45 minutes every day on social media sites such as Facebook, Twitter, Instagram, and WhatsApp. These social media platforms allow freedom of speech and there is no limit to express our views through these media. This feature enables users to post friendly and non-friendly content. The friendly contents often help us in many critical situations such as disaster management, photo sharing, video streaming, citizen engagement, and so on [4, 5, 1, 2, 3]. On the other hand, unfriendly content may leave the victim in many harmful situations that may impact on user's mental illness such as depression, sleep disorders, or even suicide in some cases [6, 7, 8]. Those unfriendly contents aim to harm a person or group based on their gender, caste, religion, wealth, and so on, and such contents are termed as hate speech or offensive contents [6, 7].

FIRE '20, Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India.
abhinavanand05@gmail.com (A. Kumar); sunil.saumya@iiitdwd.ac.in (S. Saumya); jps@nitp.ac.in (J.P. Singh)

Although offensive contents on social media are available in many forms such as textual, audio, video, and image, the majority are in the form of text. Some of the works [8, 9, 10] reported by different researchers where they used textual social media contents to identify hate and offensive text. Kumari and Singh [8] proposed convolutional neural network-based to identify hate, offensive, and profanity from English and Hindi tweets. Mishra and Pal [9] proposed an attention-based bidirectional long-short-term-memory network to identify hate, offensive, and profanity from the English, Hindi, and German tweets. Mujadia et. al [10] proposed an ensemble-based model consisting of a support vector machine, random forest, and Adaboost classifiers to identify hate contents from the English, Hindi, and German language tweets.

The current paper aims to identify the offensive or hate contents in Indo-European languages. Especially, the paper identifies the offensive tweets which are written in either English, Hindi, or German. The datasets used in the paper are floated in *HASOC-2020 track* [11] [1]. For each language, English, Hindi, and German two tasks were given. For the *sub-task A*, participants have to perform a course-grained binary classification where every tweet was either Hate and Offensive (HOF) or Non- Hate and offensive (NOT). For the *sub-task B*, participants have to further classify Hate and Offensive (HOF) tweets into either hate speech, or offensive, or profane classes. The subjectivity of the attack is different in hate speech and offensive words. Further, the extremely offensive content is termed as profane. For all the languages and tasks, we proposed different models which are based on the Bidirectional Encoder Representations from Transformers (BERT) technique.

The rest of the paper is organized as follows; Section 2 elaborates on the proposed system architecture, this is followed by Section 3 that explains model setting and experiment results, Finally, Section 4 summarizes the paper.

2. Methodology

This section discusses the different steps for the proposed fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model. For the given tasks, the organizer provided the training and development dataset with the label information. The submitted model is then tested by the organizer with the private testing datasets. The overall data statistic for English, Hindi, and German language tweets used in this work is listed in Table 1.

The text containing emoticons and emojis were converted into the corresponding text using the available dictionary[2]. After conversion of emoticons and emojis, we fixed a maximum word length of 30 for each text to give an input for the Bidirectional Encoder Representations from Transformers (BERT) model. For English dataset, we used pre-trained *distilbert-base-uncased* BERT model. For the Hindi and German dataset, we used pre-trained *distilbert-base-multilingual-cased* BERT model. The *distilbert-base-uncased*[3] is a pre-trained model trained on

[1] HASOC provides a forum and a data challenge for multilingual research on the identification of problematic content.

[2] https://github.com/NeelShah18/emot/blob/master/emot/emo_unicode.py

[3] https://huggingface.co/transformers/pretrained_models.html

Table 1
Data statistic for the English, Hindi, and German datasets

		NOT	HOF	Total	HATE	OFFN	PRFN	Total
English	Training	1852	1856	3708	158	321	1377	1856
	Validation	391	423	814	25	82	293	400
	Testing	-	-	-	-	-	-	-
Hindi	Training	2116	847	2963	234	465	148	847
	Validation	466	197	663	56	87	27	170
	Testing	-	-	-	-	-	-	-
German	Training	1700	673	2373	146	140	387	673
	Validation	392	134	526	24	36	88	148
	Testing	-	-	-	-	-	-	-

Table 2
Results for the different dataset

Language	Model	Dataset		Task-A				Task-B		
				Precision	Recall	F_1-score		Precision	Recall	F_1-score
English	distilbert-base-uncased	Val	HOF	0.89	0.84	0.87	HATE	0.38	0.36	0.37
			NOT	0.84	0.89	0.86	OFFN	0.57	0.40	0.47
			Macro Avg.	0.87	0.87	0.86	PRFN	0.84	0.91	0.88
							Macro Avg.	0.60	0.56	0.57
		Test	Macro Avg.	-	-	0.5031	Macro Avg.	-	-	0.1623
Hindi	distilbert-base-multilingual-cased	Val	HOF	0.57	0.42	0.48	HATE	0.64	0.29	0.40
			NOT	0.78	0.86	0.82	OFFN	0.54	0.90	0.67
			Macro Avg.	0.67	0.64	0.65	PRFN	0.00	0.00	0.00
							Macro Avg.	0.39	0.39	0.36
		Test	Macro Avg.	-	-	0.5300	Macro Avg.	-	-	0.0940
German	distilbert-base-multilingual-cased	Val	HOF	0.76	0.52	0.62	HATE	0.72	0.54	0.62
			NOT	0.85	0.94	0.90	OFFN	0.52	0.31	0.39
			Macro Avg.	0.81	0.73	0.76	PRFN	0.72	0.89	0.79
							Macro Avg.	0.65	0.58	0.60
		Test	Macro Avg.	-	-	0.5109	Macro Avg.	-	-	0.1214

the Toronto Book and full English Wikipedia corpus. The *distilbert-base-uncased* contains 6 layers, 768 dimension and 12 heads with 66 Million parameters. The *distilbert-base-multilingual-cased*[4] is a multi-lingual pre-trained model trained on the 104 different languages of Wikipedia texts. The *distilbert-base-multilingual-cased* contains 6 layers, 768 dimension and 12 heads with 134 Million parameters. To fine tune the respective *distilbert-base-uncased* and *distilbert-base-multilingual-cased* models, a batch size of 32 and learning rate of $2 * 10^{-5}$ were used to train the model for the 20 epochs.

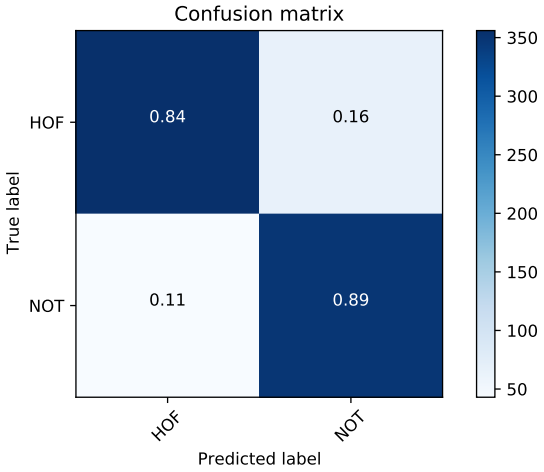

Figure 1: Confusion matrix for task-A English

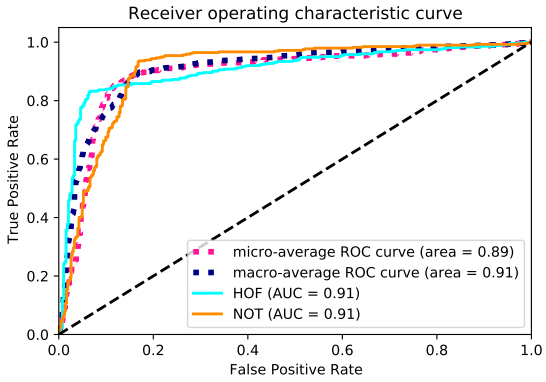

Figure 2: ROC for task-A English

3. Results

The results of the fine-tuned BERT models for the English, Hindi, and German dataset is listed in Table 2. We listed the results of the validation (Val) datasets and the results provided by the organizers for the private testing (Test) datasets. For English sub-task A, a macro averaged precision, recall, and F_1-score of 0.87, 0.87, and 0.86, respectively were achieved for the validation dataset whereas a macro F_1-score of 0.5031 was achieved for the test dataset (as can be seen in Table 2. The confusion matrix and ROC curve for English sub-task A validation dataset can be seen from Figures 1 and 2, respectively. For English sub-task B, a macro averaged precision, recall, and F_1-score of 0.60, 0.56, and 0.57 were achieved respectively for the validation dataset. For the testing dataset, a macro F_1-score of 0.1623 was achieved. The confusion matrix for English sub-task B can be seen from Figures 3 and 4, respectively for the validation dataset.

For Hindi sub-task A, a macro averaged precision, recall, and F_1-score of 0.67, 0.64, and 0.65 were achieved for the validation dataset whereas a macro averaged F_1 score of 0.5300 for the test dataset. The confusion matrix and ROC curve for the Hindi sub-task A validation dataset can be seen from Figures 5 and 6, respectively. For Hindi sub-task B, a macro averaged precision, recall, and F_1-score of 0.39, 0.39, and 0.36 were achieved for the validation dataset. The confusion matrix for Hindi sub-task B validation dataset can be seen from Figures 7 and 8, respectively.

For German sub-task A, a macro averaged precision, recall, and F_1-score of 0.81, 0.73, and 0.76 were achieved for the validation dataset whereas a macro averaged F_1-score of 0.5109 was achieved for the test dataset. The confusion matrix and ROC curve for the German sub-task A validation dataset can be seen from Figures 9 and 10, respectively. Similarly, for the German sub-task B, a macro averaged F_1-score of 0.65, 0.58, and 0.60 were achieved for the validation dataset. The confusion matrix and ROC curve for the German sub-task B validation dataset

[4]https://huggingface.co/transformers/pretrained_models.html

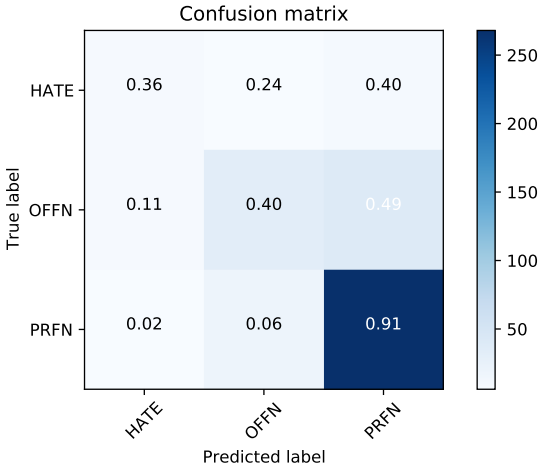

Figure 3: Confusion matrix for task-B English

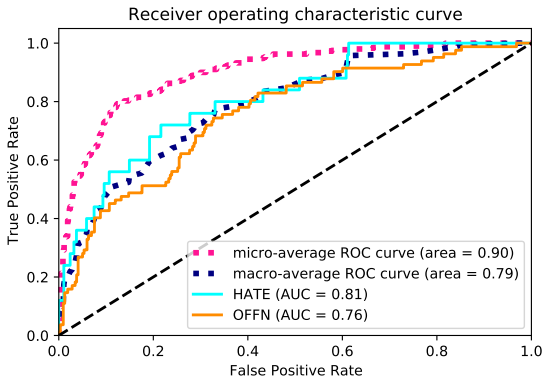

Figure 4: ROC for task-B English

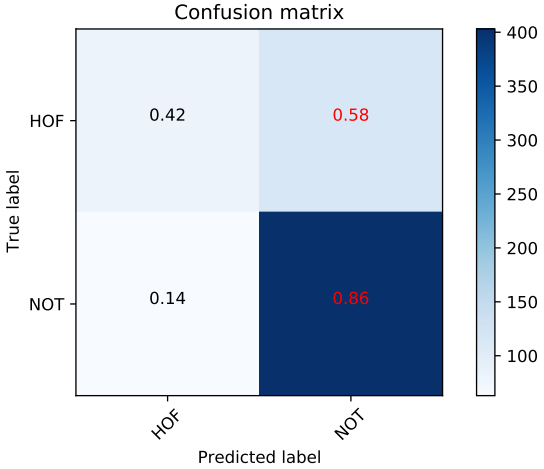

Figure 5: Confusion matrix for task-A Hindi

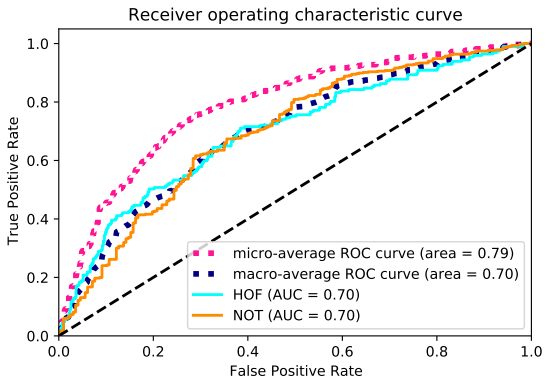

Figure 6: ROC for task-A Hindi

can be seen from Figures 11 and 12, respectively.

4. Conclusion

Hate and offensive social media contents may leave the victim in many harmful situations that may impact on user's mental illness such as depression, sleep disorders, or even suicide in some cases. Therefore identification of such hate and offensive social media contents are essential natural language processing tasks. In this work, we have proposed a fine-tuned BERT model for the identification of hate and offensive tweets from the English, Hindi, and German language text. The proposed fine Tuned BERT model achieved significant performance for all

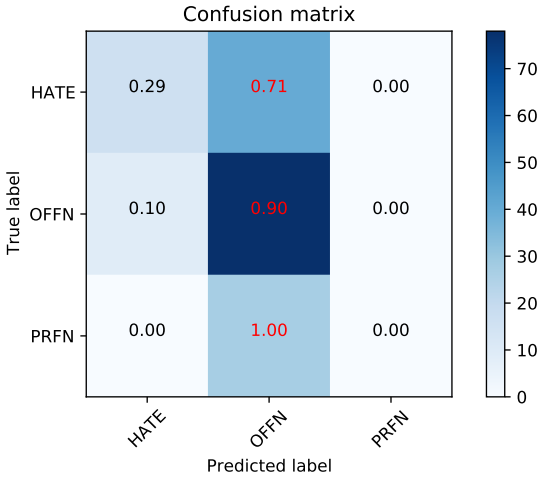

Figure 7: Confusion matrix for task-B Hindi

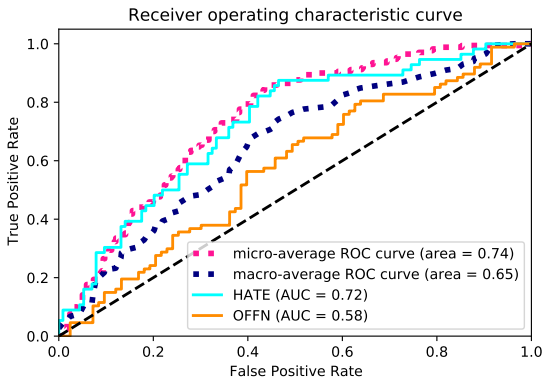

Figure 8: ROC for task-B Hindi

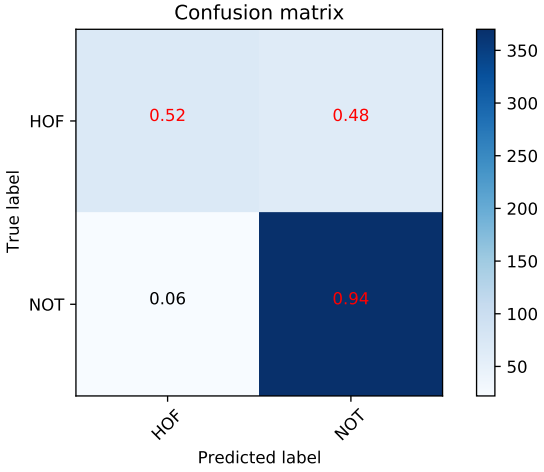

Figure 9: Confusion matrix for task-A German

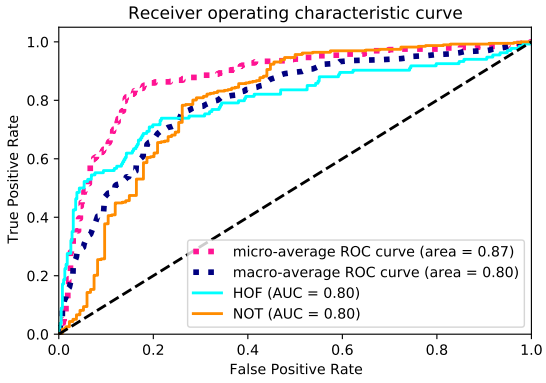

Figure 10: ROC for task-A German

three English, Hindi, and German language tweets. For English subtasks A and B F1-score reported were 0.5031 and 0.1623, for Hindi subtasks A and B F1-score reported were 0.5300 and 0.0940 and for German subtasks A and B F1-score reported were 0.5109 and 0.1214 respectively. The future work may explore the role of embedding with fine tuned BERT model for better classification performance.

References

[1] A. Kumar, N. C. Rathore, Relationship strength based access control in online social networks, in: Proceedings of First International Conference on Information and Commu-

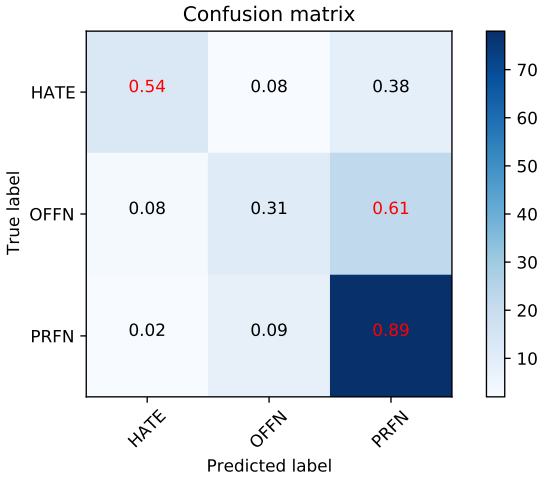

Figure 11: Confusion matrix for task-B German

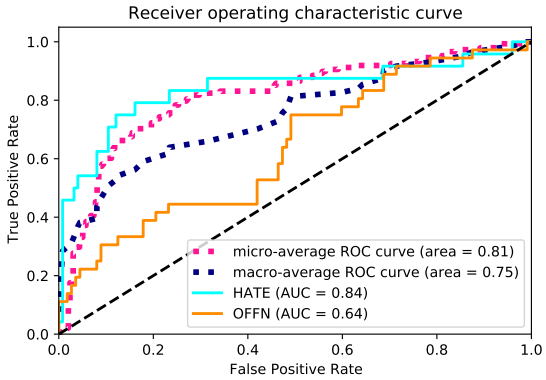

Figure 12: ROC for task-B German

nication Technology for Intelligent Systems: Volume 2, Springer, 2016, pp. 197–206.

[2] A. Kumar, J. P. Singh, Location reference identification from tweets during emergencies: A deep learning approach, International journal of disaster risk reduction 33 (2019) 365–375.

[3] A. Kumar, J. P. Singh, Y. K. Dwivedi, N. P. Rana, A deep multi-modal neural network for informative twitter content classification during emergencies, Annals of Operations Research (2020) 1–32.

[4] J. P. Singh, A. Kumar, N. P. Rana, Y. K. Dwivedi, Attention-based lstm network for rumor veracity estimation of tweets, Information Systems Frontiers (2020) 1–16. doi:https://doi.org/10.1007/s10796-020-10040-5.

[5] A. Kumar, J. P. Singh, S. Saumya, A comparative analysis of machine learning techniques for disaster-related tweet classification, in: 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129), IEEE, 2019, pp. 222–227.

[6] 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.

[7] 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, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[8] K. Kumari, J. P. Singh, Ai ml nit patna at hasoc 2019: Deep learning approach for identification of abusive content., in: FIRE (Working Notes), 2019, pp. 328–335.

[9] A. Mishra, S. Pal, Iit varanasi at hasoc 2019: Hate speech and offensive content identification in indo-european languages., in: FIRE (Working Notes), 2019, pp. 344–351.

[10] V. Mujadia, P. Mishra, D. M. Sharma, Iiit-hyderabad at hasoc 2019: Hate speech detection., in: FIRE (Working Notes), 2019, pp. 271–278.

[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.