Notebook for Emotions & Threat Detection in Urdu @ FIRE 2022

Bin Wang[1], Hui Ning[1,*]

[1]*Harbin Engineering University, 145 Nantong St, Harbin, 150000, China*

Abstract

With social media becoming an important medium for spreading information in people's daily lives, the detection of threatening speech has become increasingly necessary. Recently, FIRE 2022 has released "EmoThreat: Emotions & Threat Detection in Urdu" task which include two subtasks. We have chosen to participate in "Task B: Threatening Language Detection Task in Urdu". Several BERT models based on transformers were used to accomplish this task. As a result, we came 4th in subtask 1 and 5th in subtask 2.

Keywords

EmoThreat, Transformer, BERT

1. Introduction

Today basically anyone can communicate and share information, opportunities and ideas in real time on Internet social media, opportunities and ideas in real time. Social media has revolutionized how humans interact, providing them with unprecedented opportunities to satisfy their social needs [1]. With social media becoming an important medium for spreading information in people's daily lives, the detection of threatening speech has become increasingly necessary.

Recently, FIRE 2022 has released "EmoThreat: Emotions & Threat Detection in Urdu" task which include two subtasks [2] [3]. Urdu is the national language of Pakistan and also a widely spoken language in Indian sub-continent. In recent times, data pertaining to Urdu language is increasing tremendously on web [4]. We have chosen to participate in "Task B: Threatening Language Detection Task in Urdu", which is related to previous work in FIRE 2021 [5] [6]. This subtask aims to identify whether a tweet in Urdu is threatening or not, and if it is judged to be a threatening tweet, the task further requires to distinguish whether the tweet is a threat to an individual or a group.

*Corresponding author

FIRE 2022: Forum for Information Retrieval Evaluation, December 9–13, 2022, India

sgomw@hrbeu.edu.cn (B. Wang); ninghui@hrbeu.edu.cn (H. Ning)

0000-0002-6711-5887 (B. Wang)

2. Related work

The BERT model is now widely used in solving text classification problems. A taxonomy that classifies the state-of-the-art of tweet-based bot detection techniques was provided in [7], they also describe the shallow and deep learning techniques for tweet-based bot detection, along with their performance results. [8] explored several machine learning models such as XGboost, LGBM, m-BERT based models for abusive and threatening content detection in Urdu based on the shared task. In [9], a machine learning model is developed for detecting threats from Twitter tweets, accordingly, the Naïve Bayes classifier and flask microweb framework were used to build the model by using the python programming language. [10] approached the task in HASOC 2021 using the Transformers-based Roberta model which has shown remarkable results in NLP tasks such as sentence classification. [11] introduce and release a new dataset for threatening language detection in Urdu tweets to further research in this language. Previous study [12] presents guidelines for data annotation process and developed two linguistic resources: (i) Annotated corpus in Roman Urdu Language for cyberaggression and offensive language detection, the process of data annotation involved bilingual annotators instead of crowdsourcing, (ii) Processing textual information for NLP tasks involves Stop-word elimination as a sub phase, stop words carry least semantic information and increase feature space as compared to the other tokens and index terms in corpora. [13] create a well-balanced dataset by adding a neutral class to a benchmark dataset consisting of four emotions: fear, sadness, joy, and anger, on this extended dataset, we investigate the use of Support Vector Machine (SVM) and Bidirectional Encoder Representations from Transformers (BERT) for emotion recognition.

3. Dataset

The organizers provided a total of 3564 tweets in Urdu, of which 1782 were marked as threatening tweets, 1782 were marked as non-threatening tweets, while 1341 of the 1782 threatening tweets were marked as threatening tweets for groups and the remaining 441 tweets were marked as threatening tweets for individuals [14] [15] [16].

4. Methodology

4.1. Subtask 1

For subtask 1 of task B, we tried two different BERT-based pre-training models from the huggingface community [17]. One is called “Hate-speech-CNERG/dehatebert-mono-arabic” [18], a model used in the HASOC 2021 task of FIRE@2021, and the other model is called “Hate-speech-CNERG/urdu-abusive-MuRIL” [19], two both models are from the same open source model contributor.

For the “Hate-speech-CNERG/dehatebert-mono-arabic” model, we tried the combination of batch size = 24, epochs = 10, while for the “Hate-speech-CNERG/urdu-abusive-MuRIL” model, we tried for the “Hate-speech-CNERG/urdu-abusive-MuRIL” model, we tried the combination of batch size=24, epochs = 10 and batch size=24, epochs = 15, and we can see that the results is improving in Table 1.

Table 1
Results of subtask 1

Run	F1	Accuracy	ROC-AUC
Run1	0.6140	0.686	0.609
Run2	0.648	0.699	0.644
Run3	0.681	0.722	0.679

4.2. Subtask 2

Considering that if the labeled test set is directly classified into two categories, threatening tweets and non-threatening tweets, based on subtask 1, subtask 2 can again be considered as a binary classification task if the labeled threatening tweets are then classified to target individuals or to target groups.

We simplified subtask 2 to a binary classification task based on the prediction results we obtained in subtask 1, and fed it into the "Hate-speech-CNERG/dehatebert-mono-arabic" model as a simple experiment. As we did in subtask 1, different combinations of batch size and epochs were tried one after another, and the results were obtained according to the feedback as shown in Table 2.

Table 2
Results of subtask 2

Run	F1	Accuracy	ROC-AUC
Run1	0.378	0.656	0.564
Run2	0.41	0.666	0.588
Run3	0.387	0.656	0.564

5. Results

We ended up with F1 scores of 0.681, Accuracy scores of 0.722, and ROC-AUC scores of 0.679 for subtask 1, and F1 scores of 0.41, Accuracy scores of 0.666, and ROC-AUC scores of 0.588 for subtask 2. As shown in Table 3.

Table 3
Final result

Subtask	F1	Accuracy	ROC-AUC
Subtask1	0.681	0.722	0.679
Subtask2	0.41	0.666	0.588

We were ranked 4th in subtask 1 and a 5th in subtask 2.

6. Error Analyses

Regarding subtask 1, at first we were concerned about the occurrence of overfitting, so we chose smaller epochs. However, we can see that the accuracy improves significantly only as the epochs increase, indicating that there is still much room for parameter tuning. Regarding subtask 2, we basically just considered it as a dichotomous classification task after simply trying it out. Training it with a more suitable model may give better results.

7. Conclusions

Transformer-based BERT models have been used very widely and effectively in NLP tasks. This helps one to reduce much of the workload, and instead of building and tuning a model from scratch, we can choose to directly call an off-the-shelf model that has been proven to be effective in similar problems. We simply called the pre-trained BERT model from the huggingface community and completed the task.

8. Acknowledgments

Thanks to the providers of the open source model, to the huggingface community and to the organisers of the task, without whom we would not have been able to perform the task with ease.

References

[1] J. Chen, Y. Wang, et al., Social media use for health purposes: systematic review, Journal of medical Internet research 23 (2021) e17917.

[2] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, Overview of EmoThreat: Emotions and Threat Detection in Urdu at FIRE 2022, in: CEUR Workshop Proceedings, 2022.

[3] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, EmoThreat@FIRE2022: Shared Track on Emotions and Threat Detection in Urdu, in: Forum for Information Retrieval Evaluation, FIRE 2022, Association for Computing Machinery, New York, NY, USA, 2022.

[4] A. Khattak, M. Z. Asghar, A. Saeed, I. A. Hameed, S. Asif Hassan, S. Ahmad, A survey on sentiment analysis in urdu: A resource-poor language, Egyptian Informatics Journal 22 (2021) 53–74. URL: https://www.sciencedirect.com/science/article/pii/S1110866520301171. doi:https://doi.org/10.1016/j.eij.2020.04.003.

[5] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. Amjad, O. Vitman, A. Gelbukh, Overview of abusive and threatening language detection in urdu at fire 2021, CEUR Workshop Proceedings 3159 (2021) 744–762. Publisher Copyright: © 2021 Copyright for this paper by its authors.; null ; Conference date: 13-12-2021 Through 17-12-2021.

[6] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. I. Amjad, O. Vitman, A. Gelbukh, Overview of the shared task on threatening and abusive detection in Urdu at FIRE 2021, in: FIRE (Working Notes), CEUR Workshop Proceedings, 2021.
[7] A. Derhab, R. Alawwad, K. Dehwah, N. Tariq, F. A. Khan, J. Al-Muhtadi, Tweet-based bot detection using big data analytics, IEEE Access 9 (2021) 65988–66005. doi:10.1109/ACCESS.2021.3074953.
[8] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. I. Amjad, O. Vitman, A. Gelbukh, UrduThreat@ FIRE2021: Shared track on abusive threat identification in Urdu, in: Forum for Information Retrieval Evaluation, 2021, pp. 9–11.
[9] S. H. Sadi, M. R. H. Pk, A. M. Zeki, Threat detector for social media using text analysis, International Journal on Perceptive and Cognitive Computing 7 (2021) 113–117.
[10] S. Kalraa, M. Agrawala, Y. Sharmaa, Detection of threat records by analyzing the tweets in urdu language exploring deep learning transformer-based models (2021).
[11] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening language detection and target identification in urdu tweets, IEEE Access 9 (2021) 128302–128313.
[12] A. Dewani, M. A. Memon, S. Bhatti, Development of computational linguistic resources for automated detection of textual cyberbullying threats in roman urdu language, 3 c TIC: cuadernos de desarrollo aplicados a las TIC 10 (2021) 101–121.
[13] I.-A. Albu, S. Spînu, Emotion detection from tweets using a bert and svm ensemble model, arXiv preprint arXiv:2208.04547 (2022).
[14] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening language detection and target identification in urdu tweets, IEEE Access 9 (2021) 128302–128313. doi:10.1109/ACCESS.2021.3112500.
[15] N. Ashraf, A. Rafiq, S. Butt, H. M. F. Shehzad, G. Sidorov, A. Gelbukh, Youtube based religious hate speech and extremism detection dataset with machine learning baselines, Journal of Intelligent & Fuzzy Systems (2022) 1–9.
[16] N. Ashraf, R. Mustafa, G. Sidorov, A. Gelbukh, Individual vs. group violent threats classification in online discussions, in: Companion Proceedings of the Web Conference 2020, 2020, pp. 629–633.
[17] 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, arXiv preprint arXiv:1910.03771 (2019).
[18] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection, arXiv preprint arXiv:2004.06465 (2020).
[19] M. Das, S. Banerjee, A. Mukherjee, Data bootstrapping approaches to improve low resource abusive language detection for indic languages, arXiv preprint arXiv:2204.12543 (2022).