Detection of Threat Records by Analyzing the Tweets in Urdu Language Exploring Deep Learning Transformer - Based Models

Sakshi Kalra[a], Mehul Agrawal[a] and Yashvardhan Sharma[a]

[a]*Department of Computer Science and Information Systems, Birla Institute of Technology and Science Pilani, Pilani Campus, Rajasthan, India*

Abstract

As humans, we express sadness, anger, happiness, frustration, bullying, etc., in both physical and virtual worlds. In the virtual world, i.e., social media, we use textual ways to express ourselves. Due to the lack of offensive and threatening language detection mechanisms aggressive behavior in social media is not always followed by an immediate consequence. But the impact of these posts on the victim can cause prolonged mental illness and instigate fear for social media platforms. This paper aims to identify threatening posts using deep learning transformer-based models such as Roberta. The Urdu tweet dataset used in this study has been provided by HASOC-2021 which aims to identify Hate speech and offensive remarks without human assistance. We submitted our model in its subtask B of the 4th subtrack(Abusive and Threatening language detection in Urdu), secured 2nd position on the public leaderboard, and obtained Weighted f1 of 0.5346 and ROC AUC of 0. 8199.

Keywords

Threatning language detection, Hate speech, Label classification, Versions of BERT, HASOC

1. Introduction

With the expansion of the Internet, Social media has become a nursery of toxic and unethical content over the years. It is flooding with manipulative and hateful information leading to disharmony and violence in society. A lot of research has gone into identifying threatening and offensive language by many social media giants. Detecting hate speech using well-defined algorithms will go a long way in filtering out inappropriate language from powerful platforms used for general discourse and prevent the psychological and physical consequences of these abusive comments on its victims. HASOC 2021 is searching for technology to identify Hate speech and offensive remarks without human assistance. The challenge is divided into four subtracks. We participated in the 4th subtrack, which aimed at identifying Abusive and Threatening language in Urdu[1] and code is available in the github repository. [2]

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

p20180437@pilani.bits-pilani.ac.in (S. Kalra); f20180955@pilani.bits-pilani.ac.in (M. Agrawal); yash@pilani.bits-pilani.ac.in (Y. Sharma)

https://www.bits-pilani.ac.in/pilani/yash/profile (Y. Sharma)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://www.Urduthreat2021.cicling.org/home

[2]https://github.com/Kalra-Sakshi/HASOC-Subtask-B-files.git

This subtrack is further divided into two subtasks. Subtask A focuses on detecting Abusive language and subtask B on threatening language using Twitter tweets in Urdu. This paper aims to describe the methodology used for subtask B. There is a subtle difference between abusive and threatening language. The current definition of abusive speech is anything that directly attacks people based on what is known as their "protected characteristics" — race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity, or severe disability or disease. For example. "[Religious Group] are viruses; they are making this country sick." On the other hand, a threat is a communication intended to inflict harm or loss on another person, such as "Let's kill these cops cuz they don't do us no good / pullin' out your Glock out 'cause I live in the 'hood" and "I'ma jam this rusty knife all in his guts and chop his feet."

We approached the task using the Transformers-based Roberta [1] model which has shown remarkable results in NLP tasks such as sentence classification. The provided Urdu dataset is fine-tuned using a pre-trained RoBERTa transformer model from the HuggingFace library [2].

2. Related Work

A lot of research has already been done in identifying hate speech on social media platforms. Waseem and Hovy [2] proposed models to identify sexism and racism in hate speech using character n-grams which performed better than word n-grams. Alfina et al. [3] has proposed traditional machine learning classifier models such as Naïve Bayes, SVM, Bayesian Logistic Regression, and Random Forest Decision Tree to identify hate speech in the Indonesian language.

Kamble et al. [4] used domain-specific word-embeddings for hate speech identification in Hindi-English code-switched tweets. Authors of [5, 6, 7] proposed bidirectional encoder representations from BERT to detect hate and offensive language in English texts. [8] used a CNN-based model for the same classification task in code-switched Hindi datasets.

Chen et al. used NLP methods to propose sentence lexical and syntactic features for offensive speech detection [9]. Huang et al. integrated the textual features with social network features, which helped in improved cyberbullying detection on social media platforms[10]. NLP researchers have invested well enough in developing models to identify offensive content or hate speech on social media platforms in the last couple of years. Many NLP methods or tools are proposed to the said problem. The Hate Speech and Offensive Content Identification (HASOC 2021) has been organized to identify hateful and threatening language on social media platforms, particularly for low-resource languages like Urdu. Amjad et al. [10],[11] represent the first distributed task for fake news detection in the Urdu language. Amjad et al. in [12] presented a novel dataset for analyzing threatening and non-threatening language in the Urdu language. Amjad et al. in [13] presented a novel dataset for analyzing threatening and non-threatening language in the Urdu language. In [14] they tried automatic abusive language detection of the Urdu tweets.

3. Dataset

The subtask B in the HASOC challenge for Urdu is a binary classification task. We need to categorize the sentences in the Urdu tweet dataset into Threatening (THR) and Non-Threatening

(NON-THR) categories. There are a total of 6000 tweets in the dataset. The dataset is imbalanced, with 4929 tweets categorized as Non-Threatening (NON-THR) and 1071 as Threatening (THR). The dataset distribution of the task can be seen in fig 1.

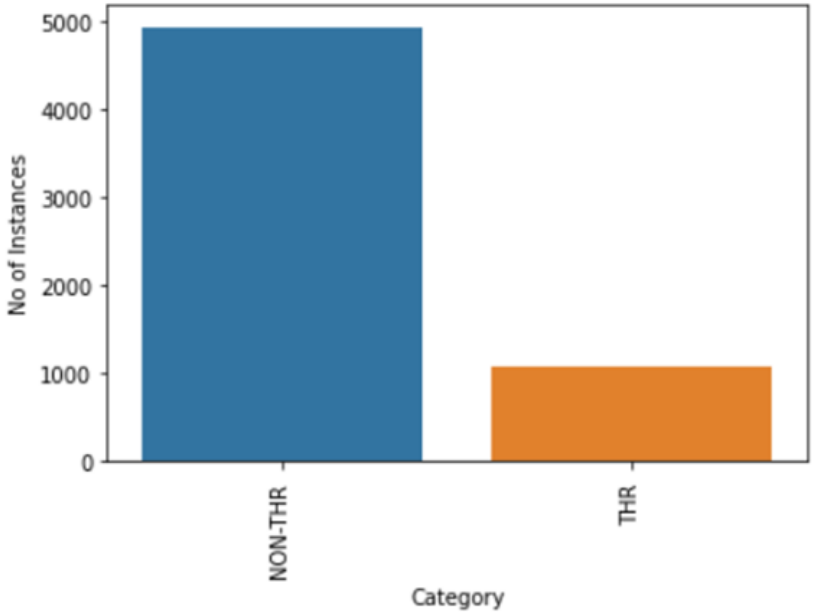

Figure 1: Dataset Distribution for Subtask B: Threatening Language Detection (HASOC 2021 Urdu)

4. Handling the Class Imbalanced Issue

One approach to address the problem of data imbalance in the training dataset is to resample it randomly. There are two methods to resample the dataset randomly: undersampling, i.e., deleting examples from the majority class, and oversampling, i.e., duplicating samples from the minority class[13] Since the number of training instances is already relatively more minor, deleting examples from the majority class will further reduce the training instances; hence we oversampled the dataset using the imblearn[15] library. RandomOverSampler with a sampling strategy of 0.5, i.e., making the ratio of minority to majority class 0.5. To normalize Urdu text, we used the Urduhack library normalization module. It replaces Arabic characters with correct Urdu characters hence brings all the characters in the specified Unicode range (0600-06FF) for the Urdu language. It also put spaces After Urdu Punctuations and removed diacritics from Urdu text.

5. Proposed Techniques and Algorithms

Transformers-based models are state of the art in various NLP tasks such as Machine Translation, Question Answering Systems, Rumour Detection, Fake News Detection, etc. They perform

better than previous methods as they are bidirectionally trained and have a deeper understanding of the language. One of the most favorable features of these models is that they could be pre-trained on a large corpus of raw texts and later fine-tuned on downstream tasks with lesser train instances. We used a pre-trained Roberta model from the HuggingFace library, trained on Urdu news corpus in an unsupervised manner, enabling the model to develop contextual embedding representations of different words in the language. These representations can be used to initialize the weights of our model. We added a classification head that is randomly initialized for fine-tuning the model for sentence classification. We fine-tuned this model on our task (detecting Threatening language using Twitter tweets), transferring the knowledge of the pre-trained model to it (which is why doing this is called transfer learning). Figure 2 shows the flow of the proposed architecture. Figure 3 shows the fine-tuning of the Roberta model using the given training dataset and figure 4 shows the classification of a tweet from the test dataset using fine-tuned Roberta model.

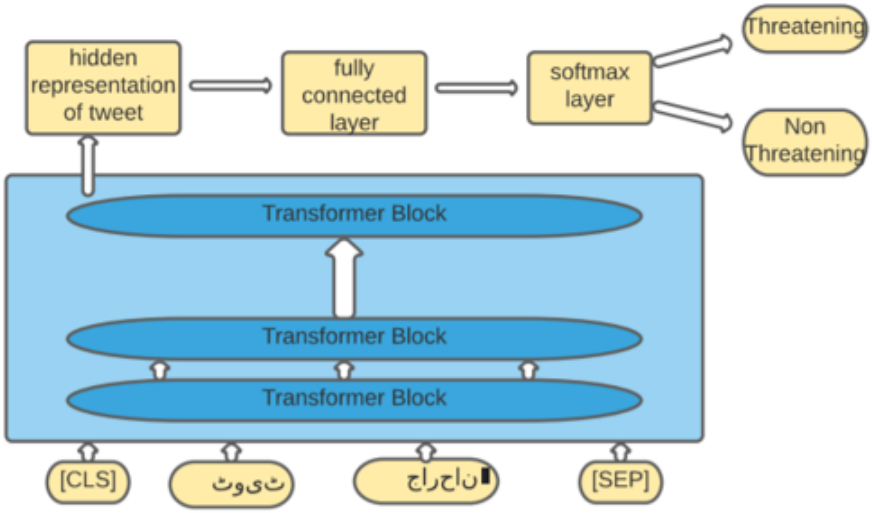

Figure 2: Proposed Architecture using the Transformer-based models for the Threatening language Detection

6. Results and Evaluations

For the proposed task, various experiments have been done on Urdu language. We used the "Urduhack/roberta-Urdu-small"model from the HuggingFace library and its corresponding tokenizer to maintain consistency with what was used during the model pre-training. The model was fine-tuned using Adam optimizer with weight decay. The maximum sequence length in each batch sent into the model was set to 256. Label smoothing cross-entropy was used for the task—a maximum learning rate of 5e-05was used for training.

The learning rates were warmed up from 0 to their maximum values and then decayed from this set maximum using a linear schedule.Table 1 lists the various hyperparameters used and

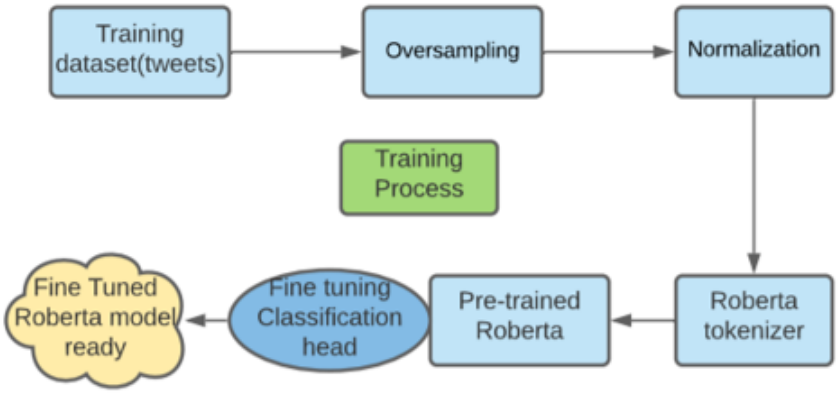

Figure 3: Fine-Tuning of the Roberta Model using the given Training Dataset

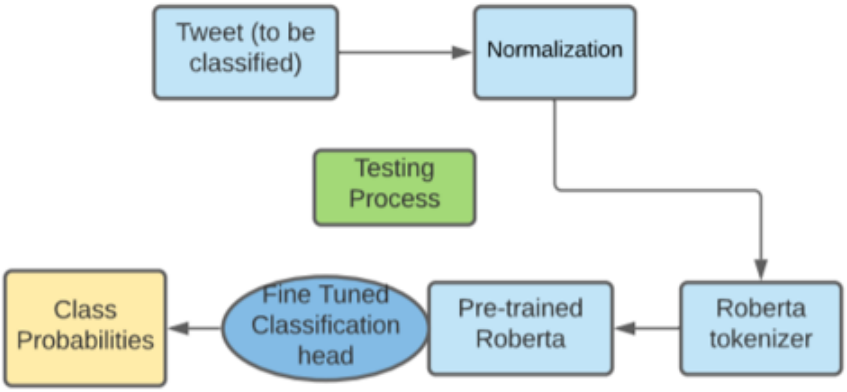

Figure 4: Classification of a Tweet from the Test Dataset using fine-tuned Roberta model

their description. Table 2 lists the model performance on the local validation set. Table 3 lists the performance of the model on the leaderboard. The organizers used weighted f1 and ROC AUC scores to evaluate the models and rank us on the leader board.

7. Conclusions and Future Work

In this paper, we presented our solution to subtask B of subtrack "HASOC - Abusive and Threatening language detection in Urdu" of HASOC 2021. We fine-tuned an Urdu Roberta model on the provided training dataset. Before training, we did some pre-processing to tackle data imbalance and normalizing Urdu text. We were placed second on the public leaderboard for the subtask. For future work, we can try with the different transformer-based architecture to get more accurate results. We can take this hate speech detection task as a multimodal aspect

Table 1
Hyperparameters used in the Task of Threatening language detection in the Urdu Language

Hyperparameter	Description
Optimizer	Adam
Learning Rate	5e-05
Number of Epochs	3
Batch Size	32

Table 2
Model Performance on the Local Validation Set(Local CV Weighted F1-score

Model	Local CV Weighted f1-score
RoBERTa + Fine-tuned Classification head	0.52

Table 3
Model Performance on the Leaderboard (Evaluation parameter= ROC AUC)

Model	Weighted f1-score	ROC AUC
RoBERTa + Fine-tuned Classification Head	0.53	0.81

by targetting both images and text and getting the visual elements for better feature extraction. We will try to extend the model by adding the multilingual aspects.

References

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[2] 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).

[3] I. Alfina, R. Mulia, M. I. Fanany, Y. Ekanata, Hate speech detection in the indonesian language: A dataset and preliminary study, in: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, 2017, pp. 233–238.

[4] S. Kamble, A. Joshi, Hate speech detection from code-mixed hindi-english tweets using deep learning models, arXiv preprint arXiv:1811.05145 (2018).

[5] H. Ahn, J. Sun, C. Y. Park, J. Seo, Nlpdove at semeval-2020 task 12: Improving offensive language detection with cross-lingual transfer, arXiv preprint arXiv:2008.01354 (2020).

[6] W. Dai, T. Yu, Z. Liu, P. Fung, Kungfupanda at semeval-2020 task 12: Bert-based multi-task learning for offensive language detection, arXiv preprint arXiv:2004.13432 (2020).

[7] M. Ibrahim, M. Torki, N. M. El-Makky, Alexu-backtranslation-tl at semeval-2020 task 12:

Improving offensive language detection using data augmentation and transfer learning, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1881–1890.
[8] 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.
[9] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, IEEE, 2012, pp. 71–80.
[10] Q. Huang, V. K. Singh, P. K. Atrey, Cyber bullying detection using social and textual analysis, in: Proceedings of the 3rd International Workshop on Socially-aware Multimedia, 2014, pp. 3–6.
[11] M. Amjad, G. Sidorov, A. Zhila, A. F. Gelbukh, P. Rosso, Overview of the shared task on fake news detection in urdu at fire 2020., in: FIRE (Working Notes), 2020, pp. 434–446.
[12] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening language detecting and threatening target identification in urdu tweets, IEEE Access (2021).
[13] S. Cateni, V. Colla, M. Vannucci, A method for resampling imbalanced datasets in binary classification tasks for real-world problems, Neurocomputing 135 (2014) 32–41.
[14] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. I. Amjad, O. Vitman, A. Gelbukh, Overview of abusive and threatening language detection in urdu at fire 2021.", in: CEUR Workshop Proceedings.(2021). CEUR Workshop Proceedings, 2021.
[15] G. Lemaître, F. Nogueira, C. K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, The Journal of Machine Learning Research 18 (2017) 559–563.

A. Online Resources

- Huggingface.