

Emotional Threat Speech Detection in Urdu Language using BERT Variants

Sakshi Kalra¹, Kushank Maheshwari¹, Saransh Goel¹ and Yashvardhan Sharma¹

¹Department of CSIS, BITS Pilani, 333031, Rajasthan, INDIA

Abstract

Threatening speech is a particular kind of content that is usually regarded as illegal and must be isolated and curbed. Threat speech identification cannot be done manually because of the volume and speed of the data being generated, i.e., over 350,000 tweets are sent per minute. Numerous studies have been done on detecting threat speech in European languages to solve this problem, but South Asian languages with limited resources have received less attention, leaving millions of users vulnerable on social media. Around 230 million people speak Urdu as their first language worldwide. This corpus of tweets is divided into three categories: Non-Threatening, Group (targeting a group), and Individual (targeting an individual). In our approach, we have fine-tuned five different pre-trained BERT models, which are transformer-based machine learning techniques. The results show that MuRIL outperformed all other models, by achieving an F1 score of 71.6%, an accuracy of 73.8% and a ROC-AUC value of 72.9% on test data.

Keywords

Threat Speech, Social Media, BERT, MuRIL, Transformers model, Multi-Class Classification

1. Introduction

Online social media platforms have exploded in popularity over the past ten years, and their user bases are expanding at an exponential rate. Users of these platforms have the freedom to share their thoughts and the opportunity to communicate with others from various groups. However, it is also used to spread, incite, promote, or justify hatred, violence, and discrimination against users based on their gender, religion, race, affiliation with particular groups, and views related to certain events or subjects (such as politics). On the one hand, this has led to exchanges of ideas and fostered relationships. On the other hand, however, it is exploited to spread hateful, offensive, derogatory, or obscene language against individuals and groups. Over 400 languages are listed in the SIL Ethnologue as being spoken in India; 24 of these languages have more than a million native speakers, while 114 have more than 10,000. Thus, there is a need for automated monitoring of threat detection.

Firms are investing heavily and advancing research in this area of threat speech detection by establishing assignments and seminars, online forums, social media enterprises, and technology. One such group is FIRE, which has been actively putting on the EmoThreat challenge to address

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

✉ p20180437@pilani.bits-pilani.ac.in (S. Kalra); f20180679@pilani.bits-pilani.ac.in (K. Maheshwari); f20190988@pilani.bits-pilani.ac.in (S. Goel); yash@pilani.bits-pilani.ac.in (Y. Sharma)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

the problem. EmoThreat 2022 is looking for ways to detect threats in speech without human intervention. The competition is broken up into two subtasks. And this paper tackles Task B. This paper contains details regarding emotions and threat detection in Urdu. This is a multi-class classification task in which the aim is to classify a tweet by a user as either non-threatening, group (targeting a group), or individual (targeting an individual).

We tackled the problem by using five different transformer-based models, namely, UrduHack, MuRIL, Multilingual-BERT, bert-base-uncased, and distilroberta. These models have displayed good outcomes in natural language processing tasks like text classification in the past, better than conventional machine learning algorithms. The urdu dataset provided by FIRE was fine-tuned using the above pre-trained transformer model from the HuggingFace library¹. The code is available from the github repository².

2. Related Work

Several researchers have already participated in the hate speech detection tasks [1], [2], [3], [4], [5],[6],[7]. Several machine learning and deep learning algorithms have been tested for automatically detecting offensive and threat speech[8]. Techniques like TF-IDF weightings and word embedding are employed in [9] and are fed into machine learning algorithms like logistic regression, random forest, and support vector classifier. Both ML models and Transformer-based models have been used for the Urdu language in [10]. According to Fire2021[11], BERT models for the identification of hate speech in the Urdu language have also been used.

Deep learning techniques[12] are currently growing in acceptance in a variety of disciplines, including language modelling, sentiment analysis, machine translation, and text classification. These include long short-term memories (LSTMs)[13], convolutional neural networks (CNNs)[14], recurrent neural networks (RNNs)[15], bidirectional encoder representations (BERT)[16]. The paper [17] lists the performance of BERT across different active learning strategies in multi-class text classification. Thus, it indicates the usage of BERT for multi-class classification involving applications in the pickup and delivery service. Another move in this direction is by [18], which compares BERT against traditional machine learning text classification. Various versions have been developed for BERT depending on its application, like DocBERT [19], which is used for document classification. BERT has been proven to perform better than existing machine learning approaches.

3. Dataset

The dataset for the task is provided by the organisers of EmoThreat'22³. Task B in the EmoThreat Urdu challenge is a multi-class classification task. A statement likely to cause damage or danger is classified as "Threatening". Threatening is further divided into "Group" and "Individual". We

¹<https://huggingface.co/>

²<https://github.com/Kushank24/fknw>

³<https://sites.google.com/view/multi-label-emotionsfire-task/dataset?authuser=0>

need to categorise the sentences in the Urdu Language dataset into the following classes: Table 1 shows the data statistics based on binary label data. Table 2 shows the multiclass label data.

- **Non-Threatening** - Tweets containing this label do not contain any threatening or profane content.
- **Group** - This label indicates that this Twitter post contains threatening content for group (s).
- **Individual** - This label indicates that this Twitter post contains threatening or profane content for an individual.

Table 1
Dataset Statistics on the basis of Binary Label Data

Data	Threatening	Non-Threatening	Total Entries
Training Data	1782	1782	3564
Testing Data	308	627	935

Table 2
Dataset Statistics on the basis of Multiclass Label Data

Data	Group	Individual	Non-Threatening	Total Entries
Training Data	441	1341	1782	3564
Testing Data	253	55	627	935

As inferred from the data, the classes Threatening and Non-Threatening have the same number of entries, but the sub-division of Threatening resulting in Individual and Group have a different number of entries. A better view can be obtained from Figure 1:

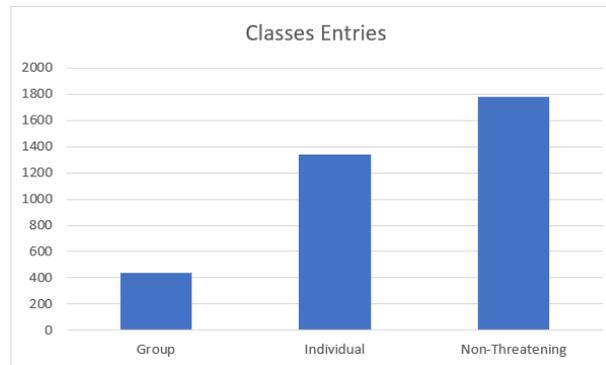


Figure 1: Training set distribution in the Urdu Dataset

4. Handling the Class Imbalanced Issue

As seen from the above figure, labels are imbalanced, so we split the data set in a stratified fashion. The proportion of data distribution in the target column is preserved by stratification,

and the train-test-split function shows the same proportion of distribution. Stratify therefore evenly distributes the target (label) throughout the training and test sets, just as it did in the original dataset. After stratification, we did oversampling of the dataset using the Imblearn library because the training instances are few and removing examples from the majority class will further reduce them. Thus, we oversampled instead of undersampling.

5. Proposed Techniques and Algorithms

For many NLP-related tasks, such as fake news identification, question answering systems, machine translation, rumour detection, etc., transformer-based models provide cutting-edge implementation. They outperform other ML methods because of their bidirectional training and improved language understanding. Pre-training is the first phase in the building of a transformer-based model, which is then fine-tuned. The model is initially trained using large language datasets (monolingual) or datasets in a variety of languages (multilingual). Only the encoder part of the transformer architecture is employed to get the word embeddings. An additional output layer is implemented to calculate the probability for classes. The various word embedding models that have been employed are listed below:

- **UrduHack**⁴ - The Urdu News Corpus was used to train Roberta-Urdu-Small. The normalisation module from urduhack was used to remove characters from other languages, such as arabic, from the training data.
- **MuRIL**⁵ - This model uses a BERT base architecture that was previously trained using corpora from 17 Indian languages from Common Crawl, Wikipedia, Dakshina, and PMINDIA.
- **bert-base**⁶ - English language pre-trained model employing masked language modelling (MLM) objective.
- **Multilingual-BERT**⁷ - This has 104 pre-trained languages. The texts are tokenized and lowercased using WordPiece, and a vocabulary with a size of 110,000 is employed. The languages with fewer resources are oversampled, whereas the languages with more Wikipedia articles are undersampled.
- **Distil-BERT**⁸ - The model has six layers, 82 million parameters, 768 dimensions, and 12 heads.

The Flowchart in Figure 2 shows the brief approach and intermediate steps.

The following Hyper-parameters were used while training the model:

- **Optimizer** - an optimizer is a function or an algorithm that modifies the attributes to reduce the overall loss and improve accuracy. In our implementation, we have used the AdamW optimizer, which is a variant of the Adam optimizer with an improved implementation of weight decay.

⁴<https://huggingface.co/urduhack/roberta-urdu-small>

⁵<https://huggingface.co/google/MuRIL-base-cased>

⁶<https://huggingface.co/bert-base-uncased>

⁷<https://huggingface.co/bert-base-multilingual-cased>

⁸<https://huggingface.co/distilroberta-base>

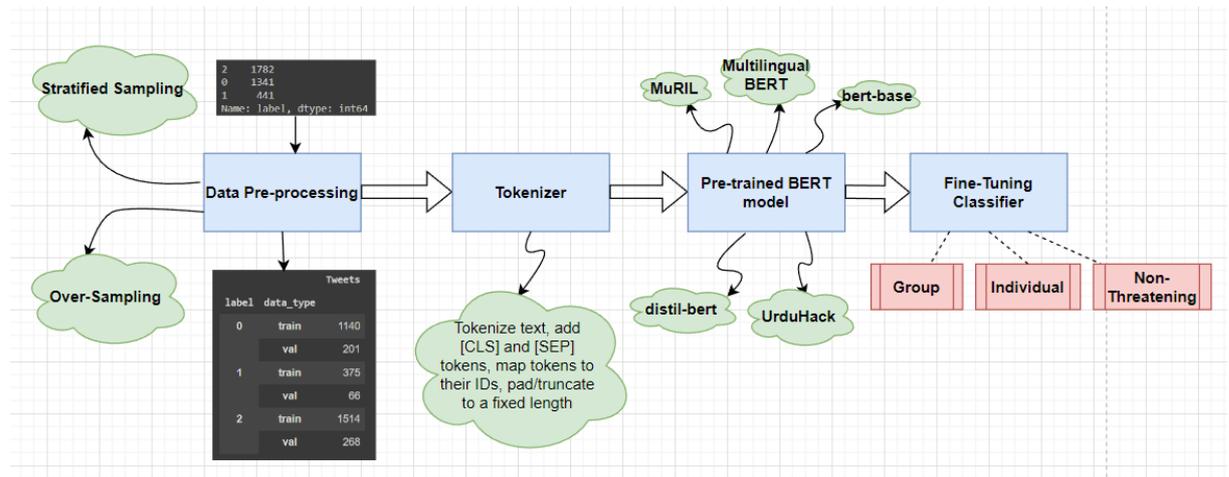


Figure 2: Flowchart of our methodology and techniques

- **Learning Rate** - an optimization technique tuning parameter that establishes the step size for each iteration. In the implementation, a learning rate of $1e-5$ is used.
- **Number of Epochs** - number of iterations over the training dataset. Five epochs were used in the implementation of the training data.
- **Batch Size** - number of samples processed before the model is updated. A batch size of 3 was used during implementation.

6. Results and Evaluations

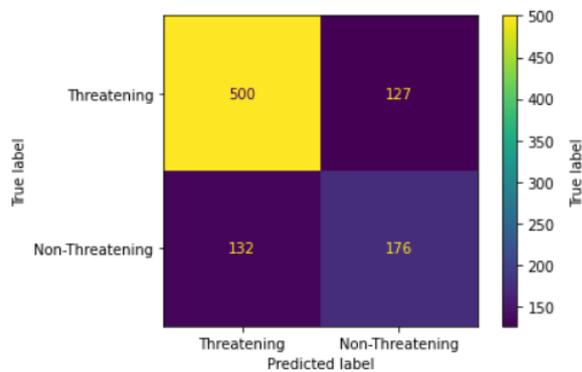
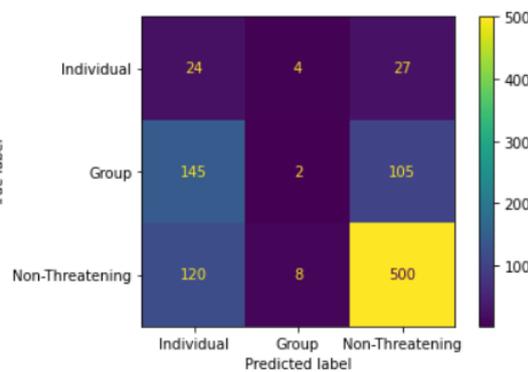
ROC-AUC, accuracy, and the F1-score are used to evaluate each model's performance. UrduHack and MuRIL gave almost similar results which were better than rest 3 BERT models. The test data provided by EmoThreat is run for the following hyperparameters: **Number of Epochs** = 5, **Batch size** = 3, **Optimizer** = AdamW, and **Learning Rate** = $1.e-5$. The results have been separately shown for both **Binary Classification** ("Threatening" vs. "Non-Threatening") and **Multi-class Classification** ("Individual" vs. "Group" vs. "Non-Threatening"). The results are shown in the below tables and figures, numbered from 5 to 14.

Table 3 shows the comparison of the five fine-tuned BERT models. As seen from Table 3 MuRIL performed best on the test data, while Multilingual BERT and UrduHack performed similarly. While distilbert and bertbase performed the worst of all models. The ROC-AUC, F1-score, and accuracy help make a complete comparison between all models. Additionally, the confusion matrix for each model also lists various errors in the classification. Finally, at last, the ROC curve for MuRIL multi-class and UrduHack multi-class is shown for the ROC value comparison. The blank values in the table show that their ROC curve was not plotted. As seen from the ROC curve, Individual vs. Rest is different in MuRIL and UrduHack, and thus UrduHack is better able to classify Individual vs. Rest as compared to MuRIL.

Table 3

Comparison of the 5 fine-tuned BERT models

Classification	Test Data Results					
	Binary Classification			Multi-Class Classification		
Metrics	Accuracy	F1	ROC-AUC	Accuracy	F1	ROC-AUC
MuRIL	73.8%	71.6%	72.9%	54.4	32.3%	60.9%
Multilingual BERT	70.37%	65.61%	65.27%	56.14%	31.11%	56.4%
UrduHack	70.2%	67.9%	69.1%	51.2%	30.1%	56.4%
Bert-base	67.37%	65.43%	67.08%	48.98%	29.45%	-
Distil-Bert	65.13	60.60%	60.62%	52.40	29.66	-

**Figure 3:** MuRIL Binary Confusion Matrix**Figure 4:** MuRIL Multi-class Confusion Matrix

7. Error Analysis

As seen from the confusion matrix, the number of false positives (FP) in the MuRIL binary class is higher than the number of FP in the mBERT binary class, while the overall accuracy for MuRIL is higher than mBERT, so for improving results, a combination of MuRIL and mBERT should be tried. Similarly, for multi-class, the false-positive total for group vs. all is lower in mBERT than in MuRIL, so a combination or an ensemble of these two would be a good model. On the other hand, the false negative for UrduHack is very low as compared to MuRIL and mBERT. Thus, if a combination of all three models or an ensemble of these three models would prove to be better

8. Conclusion and Future Work

According to the results shown above, pre-trained BERT models perform better and have a better understanding of the meaning of a sentence, making them superior learning representations. Therefore, the transfer learning strategy using pre-trained BERT models is more appropriate for identifying threat speech than standard feature extraction methods. Out of all the models,

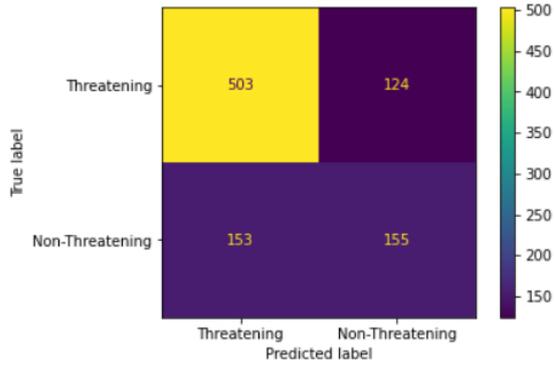


Figure 5: mBert Binary Confusion Matrix

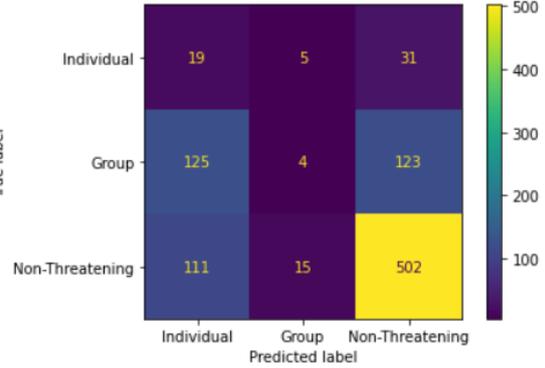


Figure 6: mBert Multi-class Confusion Matrix

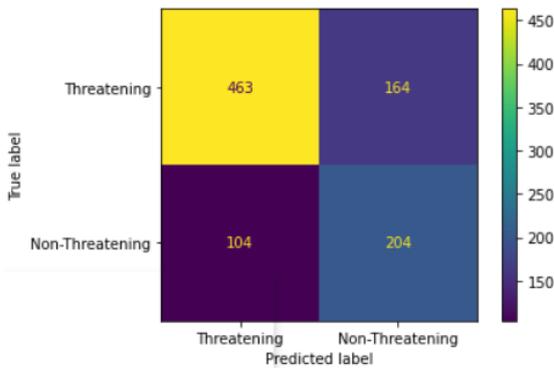


Figure 7: UrduHack Binary Confusion Matrix

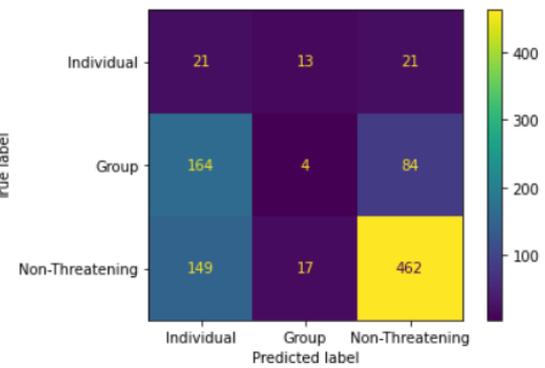


Figure 8: UrduHack Multi-class Confusion Matrix

the MuRIL performed the best. In addition, mBERT and UrduHack were comparable. We were ranked 1 on the public leaderboard. As shown by the findings above, where one model outperformed the others in a particular way, an ensemble of numerous models can also be tested to see if accuracy is increased or not. To further increase accuracy, models can be trained on a larger corpus in the future, i.e., the group and individual data points are smaller as compared to the total number of entries, thus the model is not trained well on them. The model can thus be properly trained by increasing the number of data entries. Future research on deeper transformer architectures may also be done.

References

- [1] 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.
- [2] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. I. Amjad, O. Vitman, A. Gelbukh,

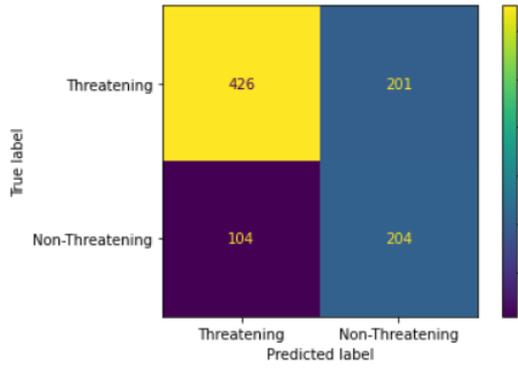


Figure 9: bertbase Binary Confusion Matrix

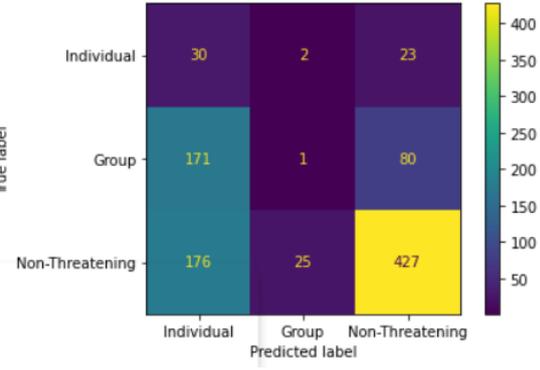


Figure 10: bertbase Multi-class Confusion Matrix

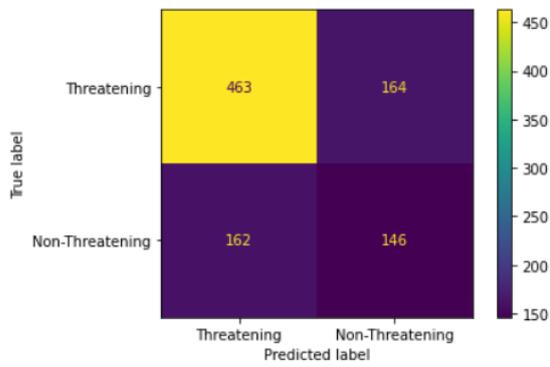


Figure 11: distilbert Binary Confusion Matrix

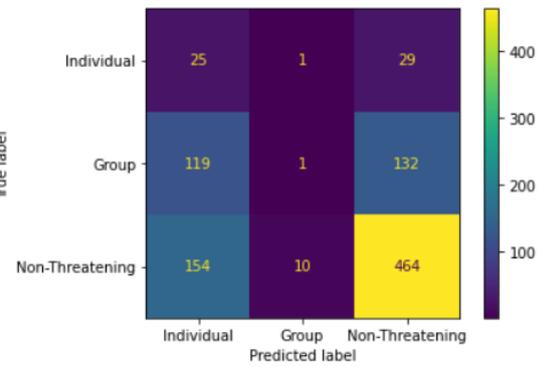


Figure 12: distilbert Multi-class Confusion Matrix

UrduThreat@ FIRE2021: Shared track on abusive threat identification in Urdu, in: Forum for Information Retrieval Evaluation, 2021, pp. 9–11.

- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, EmoTh-

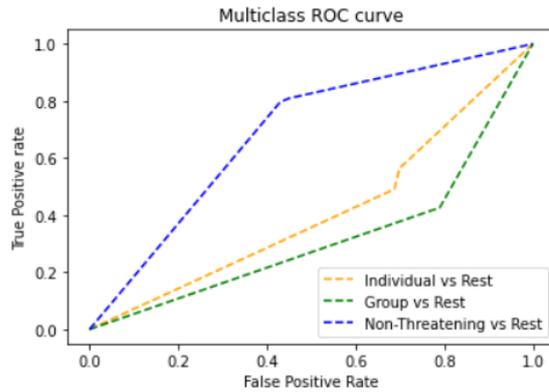


Figure 13: ROC curve for MuRIL Multiclass

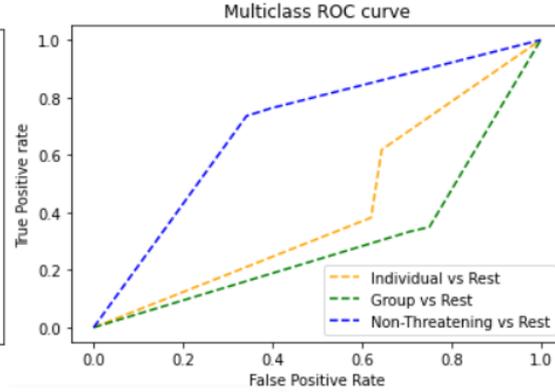


Figure 14: ROC curve for UrduHack Multi-class

reat@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.

- [8] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 512–515.
- [9] S. Kalraa, K. N. Inania, Y. Sharmaa, G. S. Chauhanb, Applying transfer learning using bert-based models for hate speech detection (2020).
- [10] S. Kalraa, Y. Bansala, Y. Sharmaa, Detection of abusive records by analyzing the tweets in urdu language exploring transformer based models (2021).
- [11] S. Kalraa, M. Agrawala, Y. Sharmaa, Detection of threat records by analyzing the tweets in urdu language exploring deep learning transformer-based models (2021).
- [12] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.
- [13] A. Bisht, A. Singh, H. Bhadauria, J. Virmani, et al., Detection of hate speech and offensive language in twitter data using lstm model, in: Recent trends in image and signal processing in computer vision, Springer, 2020, pp. 243–264.
- [14] Z. Zhang, L. Luo, Hate speech detection: A solved problem? the challenging case of long tail on twitter, Semantic Web 10 (2019) 925–945.
- [15] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Effective hate-speech detection in twitter data using recurrent neural networks, Applied Intelligence 48 (2018) 4730–4742.
- [16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [17] S. Prabhu, M. Mohamed, H. Misra, Multi-class text classification using bert-based active learning, arXiv preprint arXiv:2104.14289 (2021).
- [18] S. González-Carvajal, E. C. Garrido-Merchán, Comparing bert against traditional machine learning text classification, arXiv preprint arXiv:2005.13012 (2020).

- [19] A. Adhikari, A. Ram, R. Tang, J. Lin, Docbert: Bert for document classification, arXiv preprint arXiv:1904.08398 (2019).