# Emotions & Threat Detection in Urdu using Transformer Based Models

Anik Basu Bhaumik[1], Mithun Das[2]

[1]A.K Chowdhury School of Information Technology, University of Calcutta, Kolkata, West Bengal, India
[2]Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India

Abstract
Social media platforms have connected billions of people and helped them share their views on these platforms. However, the problem arises when malicious users abuse, show anger, and threaten others on these platforms. Therefore it is indeed necessary to detect such hostile/harmful content. So far, several studies have been conducted for hostile and negative content detection, but most of the work revolves around English. Hence to facilitate research for low-resource languages such as Urdu, the organizers of the "*EmoThreat: Emotions & Threat Detection in Urdu*"shared task at **FIRE 2022** have introduced two tasks for emotion classification and threatening language detection. In this paper, we investigate the performance of several transformer-based models and observe that the MBERT model performs the best for emotion classification. In contrast, the MURIL model performs the best for threatening tweet classification. Finally, our team hate-alert stands **3rd** in task A, **2nd** in subtask 1B and **2nd** in subtask 2B.

**Keywords**
Urdu, Threat Detection, Emotion Classification, Natural Language Processing

## 1. Introduction

Most of our population is connected to each other via the social network; the social network has and is helping us get news, express our opinion, and slowly influence our growth as a society. It has been seen that Facebook has roughly 2.93 billion monthly active users[1], Instagram has 1.21 billion monthly active users[2], and Twitter has over 450 million monthly active users globally[3]. Therefore it can understand the enormous amount of content being shared over the Internet. One of the issues with these content-sharing platforms is that occasionally bad actors share negative, abusive, threatening, and aggressive posts on this platform and endanger the well-being of millions of people [1].

To mitigate the effect of malicious content, platforms like Facebook[4] and Twitter[5] have already made guidelines that the platform users must follow to keep these platforms healthy and safe; besides, they hired moderators [2] to check the content manually. Although due

*FIRE'22: Forum for Information Retrieval Evaluation, December 9-13, 2022, India*
anikbb@gmail.com (A. B. Bhaumik); mithundas@iitkgp.ac.in (M. Das)
0000-0003-1442-312X (M. Das)


[1]https://backlinko.com/facebook-users
[2]https://www.statista.com/statistics/183585/instagram-number-of-global-users/
[3]https://www.businessofapps.com/data/twitter-statistics/
[4] https://transparency.fb.com/bn-in/policies/community-standards/hate-speech/
[5] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

to the large volume of content, it is difficult to filter all the content posted on the platforms manually. So far, several studies have been conducted to detect such negative and hostile content automatically [3, 4, 5, 6, 7], but most of the studies are centralized around the English language[8, 9].

Therefore to engage and facilitate the research around low resoruce languages, the organizers of the "*EmoThreat: Emotions & Threat Detection in Urdu* [10, 11]"[6] shared task at **FIRE 2022** have introduced two tasks for emotion classification and threat detection in Urdu. Urdu is spoken widely over South Asia; it is the official language of Pakistan. It is also widely used in regions of India and the Middle East. It has over 230 million speakers across the globe[7]. Urdu is written in Perso-Arabic script. The objective of the shared task is to devise methodologies to detect the associated emotion with a text and to classify whether a text is threatening or not.

In this paper, we investigate several transformer-based models for the classification task, which have already been seen to outperform the existing baselines and stand as a state-of-the-art model for various tasks considering hateful and abusive speech [12, 13, 14]. We conduct pre-processing, data sampling, hyper-parameter tuning, etc., to construct the model. The best models stand **3rd** in task A (Multi-label emotion classification in Urdu), **2nd** in subtask 1B( Classify the given tweet as "threatening" and "non-threatening"), and **2nd** in subtask 2B(If the tweet is classified as a "threatening" tweet, then it should be further classified as a "individual" or a "group" threat).

## 2. Related Work

Due to the exponential growth of social media platforms, sharing content on these platforms has expanded tremendously, further increasing the malicious content on these platforms. Therefore detection of such malicious content has gained significant attraction among the research community.

In 2017, Waseem et al. [5] classified abusive languages into two categories "Directed" (language directed at a specific person or thing) and "Generalized" (directed at a generalized group). Further, this category has been divided into another two categories, "Explicit" and "Implicit" (the degree to which it is explicit).

In order to accomplish the classification objective of identifying hate/offensive speech embedded in Tweets, Davidson et al. [4] provided a dataset in which thousands of tweets were categorized as "hate","offensive", and "neither". They subsequently investigated how linguistic characteristics like character and word n-grams influenced the performance of a classifier designed to identify these three categories of Tweets using this dataset. They also used features such as the number of characters, words, and syllables in each tweet, count indicators for hashtags, mentions, retweets, and URLs. The authors discovered that one of the problems with their best models was that they could not distinguish between offensive and hateful posts.

Pitsilis et al. [15] examined recurrent neural networks (RNNs) in 2018 to detect the offensive language in English. The author found that RNNs performed admirably on this task using ensemble methods, achieving an F1-score of 0.9320. RNNs preserve the outcomes of each

[6] https://sites.google.com/view/multi-label-emotionsfire-task/
[7] https://en.wikipedia.org/wiki/Urdu

step the model conducts. This technique can capture linguistic context within a text which is essential for detection. While RNNs have been projected to do well with language models, other neural network models, including CNN and LSTM, have succeeded at identifying hate/offensive speech [16, 17].

Transformer-based [18] language models, such as BERT and m-BERT [19], have recently gained popularity in various downstream tasks, like categorization and span detection. Transformer-based models have formerly been found to outperform [3] a number of deep learning models, including CNN-GRU, LSTM, and others. As a result of seeing how well these Transformer-based models function, we concentrate on developing them for our classification problem.

## 3. Dataset Description

The shared tasks present in this competition are divided into two parts. The datasets have been sampled from Twitter. The Task A is to perform multi-label emotion classification given Urdu Nastalíq tweets [20, 21, 22, 23]; it has to be classified into one or more of the following categories: *Neutral, Happiness, Surprise, Sadness, Fear, Disgust, Anger.* The task B [24, 25, 26, 27, 28] is further divided into two parts. In the first part(1B), the task is to classify a tweet as threatening or non-threatening; in the second part, the task is to classify threatening tweets into two categories: "group" or "individual" threats. The presented data has been collected and annotated from Natural Language and Text Processing Laboratory[8] at Center for Computing Research[9] of Instituto Politécnico Nacional, Mexico.

### 3.1. Task A

This task is a multi-class classification task in which tweets need to be classified into seven classes, namely: *Anger, Disgust, Fear, Sadness, Surprise, Happiness, Neutral.* The training dataset has total 7,800 instances and the test dataset has total 1,950 instances. The dataset description for this task has been represented in Table 1.

### 3.2. Task B

This is a classification task of identifying/detecting threatening language in Urdu with two sub-tasks.

- Sub-task 1B : Binary classification of the tweets as threatening and non-threatening
- Sub-task 2B : If the tweet is classified as a threatening tweet then it should be further classified as a "group" or "individual threat".

For the task B, the training dataset is having 3,564 instances and the test dataset has 935 instances which is annotated as threatening(group / individual) and non-threatening. The dataset distribution is presented in Table 2. and Table 3

[8] https://nlp.cic.ipn.mx/
[9] https://www.cic.ipn.mx/index.php/en/

| Category | Emotion classification dataset | |
|---|---|---|
| | Train | Test |
| **Neutral** | 3014 | 753 |
| **Happiness** | 1046 | 261 |
| **Surprise** | 1550 | 388 |
| **Sadness** | 2190 | 548 |
| **Fear** | 609 | 152 |
| **Disgust** | 761 | 190 |
| **Anger** | 811 | 203 |
| **Total Tweets** | 7800 | 1950 |

**Table 1**
Dataset distribution of Multi-label emotion classification (Task A)

| Category | Threat Dataset | |
|---|---|---|
| | Train | Test |
| **threatening** | 1782 | 308 |
| **non-threatening** | 1782 | 627 |
| **Total** | 3564 | 935 |

**Table 2**
Dataset distribution of threatening language detection (Task 1B)

| Category | Threat Dataset | |
|---|---|---|
| | Train | Test |
| **Group** | 1341 | 252 |
| **Individual** | 441 | 55 |
| **non-threatening** | 1782 | 628 |
| **Total** | 3564 | 935 |

**Table 3**
Dataset distribution of fine-grained threatening language detection (Task 2B)

## 4. System Description

This section explains the transformer-based models that have been explored. For task A (Multi-label emotion classification), we experimented with MBERT [19] and MURIL [29] models[10]. For subtask 1B(Binary classification of threatening language), we experimented with the following models: MBERT, MURIL, “dehatebert-mono-arabic”[11] [30] and “indic-abusive-allInOne-MuRIL”[12] [31]. The “dehatebert-mono-arabic” model is an MBERT variant, which is fine-tuned on the Arabic hate speech dataset, and the “indic-abusive-allInOne-MuRIL” model is a MURIL variant previously finetuned on eight different abusive Indic languages considering Urdu. For the sub-task 2B(fine-grained classification of threatening language), we only experimented with

[10]Code used from: https://github.com/hate-alert/IndicAbusive
[11]https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-arabic
[12]https://huggingface.co/Hate-speech-CNERG/indic-abusive-allInOne-MuRIL

| Model | Accuracy | Weighted F1 | Micro F1 | Macro F1 | Hamming loss |
|---|---|---|---|---|---|
| **MBERT** | 0.612 | 0.709 | 0.724 | 0.615 | 0.092 |
| **MURIL** | 0.519 | 0.513 | 0.610 | 0.309 | 0.117 |

**Table 4**
Multi-label Emotion Classification Results (Task A)

| Model | Accuracy | F1 Score | ROC-AUC |
|---|---|---|---|
| **MBERT** | 0.647 | 0.666 | 0.663 |
| **MURIL** | **0.716** | **0.737** | **0.729** |
| **dehatebert-mono-arabic** [30] | 0.642 | 0.687 | 0.641 |
| **indic-abusive-allInOne-MuRIL** [31] | <u>0.672</u> | <u>0.706</u> | <u>0.674</u> |

**Table 5**
Two class threatening tweet classification results (subtask 1B). The best performing model is marked in **bold** and the second best is marked in <u>underline</u>.

MBERT and MURIL models[13].

### 4.1. Multi-label Classification

The Task A is a multi-label classification problem, where each post can be classified among one or more categories. As discussed above we fine tuned transformer-based MBERT and MURIL models and added a classifier layer on top of that. BCE loss function has been used for calculating the loss.

### 4.2. Multi-class Classification

Subtasks 2A and 2B is a binary and ternary classification problems. Here we also add an extra classification layer on top of the transformer models we used. For this subtask, the Cross-Entropy loss function has been used as a loss function. Also, as seen from table 3, we can observe that the data is imbalanced; therefore, appropriate weights have been added to the classes before fine-tuning the models.

### 4.3. Tuning Parameters

The models have been run for 5 epochs with Adam optimizer[32] and initial learning rate of 2e-5. As no validation dataset was given, we divided the training data points into 85% and 15% split and used the 15% as a validation set. We predict the test set for the best validation performance.

## 5. Results

The performance of the task A, the multi-label emotion classification has been shown in Table 4. We observe that between MBERT and MURIL models, the MBERT model performs the

[13]Code used from: https://www.kaggle.com/vpkprasanna/bert-model-with-0-845-accuracy

| Model | F1 Score | Accuracy | ROC-AUC |
|---|---|---|---|
| **MBERT** | 0.473 | 0.621 | 0.626 |
| **MURIL** | 0.535 | 0.696 | 0.66 |

**Table 6**
Three class threatening tweet classification results (Task 2B)

| Actual Tweet | Translated | Actual Label | Predicted Label |
|---|---|---|---|
| بچھڑنے والوں کا کیسے نہ غم کِیا جائے یہ بَوجھ ایسا نہیں ہے کہ کم کِیا جائے میں ایک بار نہیں بار بار ہنستا ہوں کسی | How not to grieve for those who have passed away. This is not a burden to be reduced. I laugh not once, but again and again. | Sadness | Surprise |
| بڑے پاکیزہ رشتے ہوتے ہے یہ نفرت کے بدن سے کپڑے اکثر مخبت میں ہی اُترتے ہیں اشغار | There are very pure relationships, these clothes often come off from the body of hatred, Ashghar. | Sadness, Surprise | Sadness |

**Table 7**
Example of a few misclassified tweets of emotion classification

| Actual Tweet | Translated | Actual Label | Predicted Label |
|---|---|---|---|
| عوام مارتی تو مر جائے بھاڑ میں جائے بھٹو کتے کا بچا زندہ ہے | If people kill people, then they will die and go to hell. Bhutto's dog is still alive. | Non-threatening | Threatening |
| سب تحریک لبیک پاکستان کے ساتھ ملکر کفر کا مقابلہ کریں | All together with Tehreek-e-Labaik-Pakistan, fight the blasphemy. | Threatening | Non-threatening |

**Table 8**
Example of a few misclassified tweets of threat detection

best in terms of all the evaluation metrics(Acc:0.612, Weighted F1: 0.709, Macro F1:0.615). For the sub-task 1B, we observe the MURIL model perform the best(Acc: 0.716, F1:0.737, ROC-AUC:0.729) in terms of all metrics and the "indic-abusive-allInOne-MuRIL" model perform the second best(Acc: 0.672, F1:0.706, ROC-AUC:0.674). One interesting observation is that although "dehatebert-mono-arabic" and "indic-abusive-allInOne-MuRIL" models are previously finetuned on hate speech and abusive speech dataset, further fine-tuning them with the threatening tweet dataset do not outperform the vanila MURIL model. For the sub-task 2B also we obseve the MURIL model perform the best(Acc: 0.535, F1:0.696, ROC-AUC:0.66).

## 6. Error Analysis

To further understand when the model is failing, we manually inspected some misclassified tweets by the best-performing models. For the emotion classification task, we observed that the actual label itself is sometimes incorrect according to our judgment based on the translated tweets; therefore, the model is failing for such cases. For threatening tweet detection, sometimes the presence of words such as killing makes the prediction incorrect; the model cannot distinguish threatening and non-threatening tweets for such cases. We have shown the example of some misclassified tweets in Table 7 and 8.

## 7. Conclusion

In this shared task, we have experimented with several transformer-based models for multi-label emotion classification and threatening tweet detection. In specific, we explored MURIL, MBERT-based models. We observed that the MBERT model performed the best for the emotion classification, and for the threatening tweet classification, the MURIL model performed the best. Our team hate-alert stands **3rd** in task A, **2nd** in subtask 1B and **2nd** in subtask 2B.

## References


[1] J. S. Vedeler, T. Olsen, J. Eriksen, Hate speech harms: a social justice discussion of disabled norwegians' experiences, Disability & Society 34 (2019) 368–383.

[2] C. Newton, The terror queue, 2019. URL: https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video.

[3] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 14867–14875.

[4] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: ICWSM, 2017.

[5] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 78–84. URL: https://aclanthology.org/W17-3012. doi:10.18653/v1/W17-3012.

[6] M. Das, P. Saha, R. Dutt, P. Goyal, A. Mukherjee, B. Mathew, You too brutus! trapping hateful users in social media: Challenges, solutions insights, in: Proceedings of the 32nd ACM Conference on Hypertext and Social Media, HT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 79–89. URL: https://doi.org/10.1145/3465336.3475106. doi:10.1145/3465336.3475106.

[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, 2019, pp. 14–17.

[8] M. Das, B. Mathew, P. Saha, P. Goyal, A. Mukherjee, Hate speech in online social media, ACM SIGWEB Newsletter (2020) 1–8. doi:10.1145/3427478.3427482.

[9] M. Amjad, N. Ashraf, G. Sidorov, A. Zhila, L. Chanona-Hernandez, A. Gelbukh, Automatic abusive language detection in urdu tweets, ACTA POLYTECHNICA HUNGARICA (2021).

[10] 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.

[11] 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.
[12] S. Banerjee, M. Sarkar, N. Agrawal, P. Saha, M. Das, Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages, arXiv preprint arXiv:2111.13974 (2021).
[13] M. Das, S. Banerjee, P. Saha, Abusive and threatening language detection in urdu using boosting based and bert based models: A comparative approach, arXiv preprint arXiv:2111.14830 (2021).
[14] M. Das, S. Banerjee, A. Mukherjee, hate-alert@ dravidianlangtech-acl2022: Ensembling multi-modalities for tamil trollmeme classification, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, 2022, pp. 51–57.
[15] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Detecting offensive language in tweets using deep learning, ArXiv abs/1801.04433 (2018).
[16] Y. Goldberg, A primer on neural network models for natural language processing, Journal of Artificial Intelligence Research 57 (2015). doi:10.1613/jair.4992.
[17] G. L. De la Pena Sarracén, R. G. Pons, C. E. M. Cuza, P. Rosso, Hate speech detection using attention-based lstm, EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018) 235.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
[20] N. Ashraf, L. Khan, S. Butt, H.-T. Chang, G. Sidorov, A. Gelbukh, Multi-label emotion classification of urdu tweets, PeerJ Computer Science 8 (2022) e896.
[21] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, A. Gelbukh, Urdu sentiment analysis with deep learning methods, IEEE Access 9 (2021) 97803–97812.
[22] I. Ameer, N. Ashraf, G. Sidorov, H. Gómez Adorno, Multi-label emotion classification using content-based features in twitter, Computación y Sistemas 24 (2020) 1159–1164.
[23] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, Multi-class sentiment analysis of urdu text using multilingual bert, Scientific Reports 12 (2022) 1–17.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, et al., Muril: Multilingual representations for indian languages, arXiv preprint arXiv:2103.10730 (2021).
[30] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, A deep dive into multilingual hate speech classification, in: Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V, Springer International Publishing, 2021, pp. 423–439.
[31] M. Das, S. Banerjee, A. Mukherjee, Data bootstrapping approaches to improve low resource abusive language detection for indic languages, in: Proceedings of the 33rd ACM Conference on Hypertext and Social Media, HT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 32–42. URL: https://doi.org/10.1145/3511095.3531277. doi:10.1145/3511095.3531277.
[32] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.