Overview of EmoThreat: Emotions and Threat Detection in Urdu at FIRE 2022

Sabur Butt[a], Maaz Amjad[d], Fazlourrahman Balouchzahi[a], Noman Ashraf[b], Rajesh Sharma[c], Grigori Sidorov[a] and Alexander Gelbukh[a]

[a]*Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico*
[b]*Dana-Farber Cancer Institute, Harvard Medical School, United States*
[c]*Institute of Computer Science, University of Tartu, Estonia*
[d]*The University of Texas at Austin, Texas, United States*

Abstract

Emotion and targeted abuse detection i.e threat, are problems that have been studied in many rich resource languages. However, when it comes to low-resource languages such as Urdu, we find a dearth of resources and methodologies. Our paper presents the findings of the shared task "EmoThreat: Emotions and Threat detection in Urdu", where we focused on presenting resources for multi-label emotion classification (Task A) and binary threat detection (Task B) in Urdu. Task B was further divided into group and individual threat detection, making it a multi-class problem. The paper presents a summary of the methodologies and findings of the ten different participating teams. Each team also presented a thorough error analysis for the best model. The best performing system in Task A achieved a macro-F1 score of 0.687, whereas, Task B subtask 1 and subtask 2 achieved 0.716 and 0.539 macro-F1 scores respectively.

Keywords

Natural Language Processing, Emotion Detection, Threatening language Detection, Group Threats, Individual Threats, Urdu language

1. Introduction

Every language has different expressions, syntax, lexicon and vocabulary. Apart from that, languages go through a continuous process of evolution. This stands true for Urdu more than any other language. We know from the history of Urdu, that its foundation is manifested in evolution. Urdu is an amalgamation of many languages i.e. Turkish, Sanskrit, Arabic and Persian and continues to absorb words from languages that influence the demographics i.e. English. Although Urdu is written in a different script (Nastalíq) than Hindi (Devanagari), they have similar grammar and phonology. Urdu's structural similarity makes it resourceful for South Asian languages that share the same structure [1]. Due to the mixture of languages, its morphology, orthography and script Urdu requires more careful pre-processing and becomes a challenging language for Natural language processing (NLP) tasks.

FIRE 22: Forum for Information Retrieval Evaluation, December 9–13, 2022, India
sbutt2021@cic.ipn.mx (S. Butt); h.maazamjad@gmail.com (M. Amjad); fbalouchzahi2021@cic.ipn.mx (F. Balouchzahi); nomanashraf712@gmail.com (N. Ashraf); rajesh.sharma@ut.ee (R. Sharma); sidorov@cic.ipn.mx (G. Sidorov); gelbukh@gelbukh.com (A. Gelbukh)

Although Urdu has more than 230 million speakers worldwide [2], it is still a low-resource language and needs attention in multiple facets of NLP. Emotion and Threat Detection in Urdu (EmoThreat) [3] is our earnest attempt to create resources and techniques for the Urdu language that may provide assistance in understanding human behaviour online. Our first attempt is towards Emotion detection which is the heart of many NLP applications and aids in the understanding of critical semantic problems i.e. detection of irony [4], hate [5], rumours [6], threats [7, 8], sexism [9] etc. The world acknowledges the problems that comes with the growth of social media platforms [10], and we see the effect of their misuse become more impactful. In particular, numerous posts contain threatening language towards certain users and hence worsen users' experience from communication via such platforms and potentially put platform users in danger. Therefore our second attempt is for creating automatic methods for detecting threats in the Urdu language that can be helpful to avoid violence and outrageous consequences.

The paper attempts to spread awareness and encourage the community to propose more efficient methods for automated detection of multi-label emotion detection in short texts in Urdu. We highlight the collection and annotation of the first and largest datasets for detecting emotions and targeted threat in the Urdu language described in Section 2. Section 5 provides us with the baseline results for each task. Lastly, an overview and discussion of the submitted solutions for emotion and threat detection in Urdu are given in Sections 6 and 7.

2. Task Description

2.1. Task A: Multi-label Emotion Detection

We created a Nastalíq Urdu script dataset for multi-label emotion classification consisting of tweets and reviews using Ekman's six basic emotions [11] and "Neutral" sentences. The task requires you to classify the tweet as one, or more of the six basic emotions which is the best representation of the emotion or to identify instances void of emotions.

- **Anger:** also includes annoyance and rage and can be categorized as a response to a deliberate attempt of anticipated danger, hurt or incitement.
- **Disgust:** in the text is an innate response of dis-likeness, loathing or rejection to contagiousness.
- **Fear:** also including anxiety, panic and horror is an emotion in a text which can be seen triggered through a potentially cumbersome situation or danger.
- **Sadness:** also including pensiveness and grief is triggered through hardship, anguish, feeling of loss, and helplessness.
- **Surprise:** also including distraction and amazement is an emotion which is prompted by an unexpected occurrence.
- **Happiness:** also includes contentment, pride, gratitude and joy is an emotion which is seen as a response to well-being, a sense of achievement, satisfaction, and pleasure.
- **Neutral:** void of emotional affect.

2.1.1. Dataset Collection and Annotation

The dataset for Task A was taken from two separate sources. The Ekman's emotions were published and publicly presented in [12]. The benchmark dataset used Twitter hashtags for extracting relevant tweets of a particular emotion. However, since the task was to identify multiple emotions, the keywords alone were not reliable for annotation. Hence, detailed data annotation standards were formalised for expert annotators to follow and maintain consistency throughout the task. The detailed description can be found in the paper [12]. The neutral texts were collected from the paper [13] where a dataset consisting of Nastalíq Urdu texts was presented.

2.2. Task B: Threatening Language Detection

This task[1] was aimed to detect threat speech using Twitter tweets in the Urdu language without human intervention. The task was divided into two subtasks. Subtask 1 is a binary-class classification task in which participating systems are required to classify tweets into two classes, namely: (i) Threatening, and (ii) Non-Threatening.

- **Threatening** - this Twitter post contains any threatening content.
- **Non-Threatening** - this Twitter post does not contain any threatening or profane content.

Once the tweet is classified as "Threatening", then subtask 2 requires further classification of the threat into two classes: (i) Group, and (ii) Individual.

- **Group** - This Twitter post contains threatening content for targeting a group (s).
- **Individual** - This Twitter post contains threatening or profane content for threatening an individual.

2.2.1. Dataset Collection and Pre-processing

To collect the dataset for Task B, we created a dictionary of the most commonly used threatening words in Urdu. Then, we used those words as keywords on Twitter to extract tweets containing more threatening words in Urdu, which we manually added to our dictionary of threatening words. This dictionary is publicly available for research purposes.[2] Then, we used these seed words to further crawl tweets through the Twitter Developer Application Programming Interface (API)[3] using Tweepy library. We collected tweets containing any of these keywords from our dictionary for a 20-month period from January 1st, 2018 to August 31st, 2022. At this time the general elections were being held in Pakistan in July 2018. Typically, during the election season, people tend to be more expressive when supporting as well as opposing political parties. In total, we crawled 70,000 tweets containing the seed words.

Since Urdu shared many common words in Persian, Turkish and Arabic, so when we crawled tweets using our initially collected words, the Twitter API also crawled many non-Urdu tweets.

[1] https://sites.google.com/view/multi-label-emotionsfire-task/home/task-b
[2] https://github.com/MaazAmjad/Threatening_Dataset
[3] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

Since this research was primarily focused on the Urdu language, we discarded all the non-Urdu tweets manually. The threatening dataset [4] contains 9,950 tweets, 1,782 threatening tweets and the remaining tweets are non-threatening. In this shared task, to make the dataset balanced, we only randomly chose 1,782 non-threatening tweets from the remaining non-threatening tweets collected in this dataset.

It is important to mention that we created a new test set, that contains 935 tweets in total: 628 non-threatening tweets, and 307 threatening tweets. The threatening tweets are further classified into 56 tweets that target individuals, and 251 tweets that target group(s).

2.2.2. Threatening Dataset Annotation

We defined guidelines to annotate abusive and threatening tweets. To annotate the dataset annotators were recruited. All of them satisfied the following criteria: (i) country of origin-Pakistan; (ii) native speakers of Urdu; (iii) were familiar with Twitter; (iv) aged 20–35 years; (v) detached from any political party or organization; (vi) had prior experience of annotating data; (vii) educational level was a masters degree or above. We computed Inter-Annotator Agreement (IAA) using Cohen's Kappa coefficient (71%) as it is a statistical measure to check the reliability between two annotators. We provided instructions with task definitions. Hierarchical annotation schema was used and the main dataset was divided into two different datasets to distinguish between whether the language is threatening or non-threatening, abusive or non-abusive. We followed Twitter's definition to describe abusive and threatening[5] comments towards an individual or group to harass, intimidate, or silence someone else's voice.

Similarly, the task offered a dataset of tweets in Urdu annotated as threatening or non-threatening split into the training and testing parts, with the annotations of the testing part hidden from the participants. The annotation procedure for the sub-task 2 dataset followed Twitter's definition of threatening tweets[6] as those that are against an individual or group meant to threaten with violent acts, to kill, inflict serious physical harm, to intimidate, or to use violent language. The task and the evaluation procedure were identical to sub-task 2.

Table 1 and 2 show the dataset distribution of both tasks. The dataset for both tasks is publicly available on the EmoThreat website [7].

3. Literature Review

Fine-grained emotions have been extensively studied in rich resource languages such as English [14] where the emotions have been classified in Ekman's (fear, anger, joy, sadness, disgust and surprise) [11] or Plutchik's (anger, anticipation, joy, trust, fear, surprise, sadness and disgust) [15] distribution of emotions. Similarly, threatening language detection in social media texts has been one of the most crucial phenomena encompassing behaviour and emotions i.e. sexism [9], hate speech [16], abuse [8], [17], etc. but, with the element of threat in it.

[4] https://github.com/MaazAmjad/Threatening_Dataset
[5] https://help.twitter.com/en/rules-and-policies/glorification-of-violence
[6] https://help.twitter.com/en/rules-and-policies/violent-threats-glorification
[7] https://sites.google.com/view/multi-label-emotionsfire-task/dataset

Among the notable works of sentiment classification in Urdu, we see [13] a dataset comprised of reviews in Nastalíq Urdu. The reviews contained information about politics, movies, TV series and sports etc. The sentences were divided into two labels (4,758 Positive and 4,843 Negative instances) and trained on multiple machine learning and deep learning algorithms. The study revealed that Logistic Regression (LR) with word n-grams (1-3) produced the best results with 82.05% F1-score. The dataset lacked neutral tweets which were later presented in [18] for the multi-class sentiment analyses task. In the newly formed dataset, 9,312 reviews were distributed among three classes: positive, negative and neutral. The dataset was evaluated with transformers, machine learning and deep learning techniques, where the mBERT model with BERT pre-trained word embeddings outperformed all and achieved an F1 score of 81.49%. Multi-class nature of these datasets does not allow the detection of texts where multiple emotions are being expressed simultaneously. Multi-label emotion classification becomes even harder when done on the short informal text. In Urdu, multi-label emotion classification has been explored in the code-mixed setting [19]. English and Roman Urdu are widely used sometimes on social media and in text messages. Roman Urdu, can also be used for studies in Roman Hindi because of its similarity. A large dataset comprising 11,914 code-mixed (English and Roman Urdu) SMS messages was introduced [19] with a set of 12 emotions, including disgust, anger, fear, sadness, pessimism, anticipation, joy, love, optimism, surprise, trust and neutral (no emotion). The study also experimented with many different deep learning and machine learning techniques where they found a combination of OVR multi-label and SVC single-label machine learning algorithms with word uni-gram to be the most effective.

Threatening language detection in Urdu was first introduced in [20], where the authors presented a benchmark dataset in Nastaliq Urdu. The threatening language was divided to classify threats, as well as the threat target. The dataset initially consisted of 3,564 tweets, equally divided into threatening and non-threatening tweets. The authors presented detailed baselines on word and character n-grams and fastText embeddings. Support Vector Machine (SVM) classifier using fastText pre-trained word embedding obtained the best results for the target identification task, whereas, the Multilayer Perceptron (MLP) classifier with the combination of word n-gram features outperformed other classifiers in detecting threatening tweets. The dataset was then extended in the "Abusive and Threatening Language Detection Task in Urdu" [7, 8] at FIRE 2021 with 6,000 tweets in the training set and 3,950 tweets in the test set. The dataset was later extended to add group and individual threats as explained in Section 2.2.1.

Table 1
Distribution of emotions in the dataset for Task A

Emotions	Train	Test
Anger	811	203
Disgust	761	190
Fear	609	152
Sadness	2190	548
Surprise	1550	388
Happiness	1046	261
Neutral	3014	753
Total	7800	1950

Table 2
Distribution of threatening tweets in the dataset for Task B

	Threatening		Non-Threatening	Total
	Group Threat	**Individual Threat**		
Train	1341	441	1782	3564
Test	251	56	628	935

4. Evaluation Metrics

Task A used multi-label accuracy, micro-averaged F1, weighted F1 and macro-averaged F1. The ground truth annotations were used to compare the labels predicted by the participants' classifiers. All the participating teams were allowed to submit up to 3 different runs, i.e labels for the testing set generated by their proposed classifiers. The rankings were made on the basis of macro-F1 scores. We also gave the Hamming loss scores that computes the average of incorrect labels of an instance. Lower the value, the higher the performance of the classifier as this is a loss function. For Task B, we presented accuracy, macro-F1 score and ROC-AUC score. The ROC-AUC score gives an estimate of the overall quality of the model at the various level of predicted confidence thresholds and serves as a more holistic evaluator. The rankings were made on F1 scores.

5. Baselines

To get our baseline results, we used four different types of features, including character n-grams, word n-grams, stylometric features and pre-trained word embeddings. Count-based features are character n-grams and word n-grams and we used uni-, bi-, and trigrams for word n-grams while trigrams to nine grams were selected for character n-grams. Moreover, we applied Term frequency-inverse document frequency (TF-IDF) technique to choose the best count-based features. Stylometric based features contains 47 character-based, 11 word-based and 6 vocabulary based features. For the pre-trained word embeddings, we used fastText [8] library to extract 300 dimensional vectors since it provides embeddings for the Urdu language.

We approached this problem as a supervised classification task and our aim was to predict multiple emotions from the six primary emotions. To test the performance of our dataset, we utilized various machine- and deep-learning algorithms such as RF, J48, DT, SMO, AdaBoostM1, Bagging, 1D CNN, and LSTM. MEKA [9] software was used to calculate the baseline results for the machine learning algorithms while Keras [10] library was used to implement deep learning models. We used multi-label accuracy, micro-averaged F1, macro-averaged F1 and Hamming Loss (HL) for our model evaluation. Uni-gram features yield the best results on RF with the combination of a Binary Relevance (BR) transformation method, achieving 51.20% accuracy, 19.40% hamming loss, 60.20% micro-F1 and 56.10% of macro-F1 scores. The 2nd best results were achieved using 1-dimensional convolutional neural network (1D CNN) obtaining 45.00% accuracy, 36.00%

[8] https://fasttext.cc/docs/en/crawl-vectors.html
[9] http://meka.sourceforge.net
[10] https://keras.io/

hamming loss, 35.00% micro-F1 and 54.00% of macro-F1 scores. In the EmoThreat Task A, we have added neutral emotional tweets to make Task A unique and challenging. These neutral tweets are collected from [13] research article. In the train set, we have 3,014 neutral tweets while in the test set we have only 753 tweets. We applied word n-gram features and the Random Forest (RF) model to calculate the baseline results on Task A. We achieved an accuracy of 0.4835%, hamming loss of 0.1281%, micro-F1 of 0.5632% and macro-F1 of 0.3231%.

For Task B, we used Logistic Regression, Support Vector Machine, Adaboost, Decision Tree, Random Forest, and Naive Bayes with word n-grams (1-3) as features. In Task B subtask 1 we achieved the best results (0.5117 macro-F1, 0.6760 ROC-AUC) with Random Forest, and in subtask B we got the best results (0.42966 macro-F1, 0.5 ROC-AUC) with Logistic Regression using word unigrams.

6. Overview of the Submitted Approaches

This section gives a brief summary of the methods applied in the competition by the participating teams. In total 35 teams registered for the competition. Teams were allowed to participate in both or just one of the given tasks. A total of 10 different teams submitted the experimental results with the distribution of 7 teams in task A and 7 teams in Task B. We report the findings of 9 teams who submitted their methodologies in the form of technical report papers.

1. **FOSU NLP team:** The best-performing model for Task A was presented by team FOSU which used transformers. They used XLM-RoBERTa along with adamW optimizer and "ReduceLROnPlateau", a Keras module that reduces the learning rate when a metric has stopped improving.
2. **UMUTeam:** UMUteam presented a model for Task A that ranked second in the competition. The team used language-independent linguistic features, non-contextual sentence embeddings from fastText, and multilingual contextual embeddings from BERT and RoBERTa. They conducted multiple experiments leveraging Knowledge Integration (KI) and Ensemble Learning (EL). The best-performing method used a multi-input deep-learning model with Knowledge Integration (KI) of all the above-mentioned features.
3. **ERTIM:** Team ERTIM presented three distinct methods. The first method comprised transformer models (BERT-large, MuRIL, ALBERT) taken from HuggingFace [11]. In Task A, the participants analyzed the misclassified tweets and created an external dataset with similar tweets to enhance the model. The second technique comprised of Unsupervised multiword expressions (MWE), where the team first extracted autonomous chunks of text using unsupervised text segmentation. Those tokens (character and word) were then used as features for logistic regression to create a seven versus one multi-output design. Lastly, a linguistic approach was used in Task B using lexical and syntactic cues fed to machine learning models (Logistic Regression, LinearSVC, Stochastic Gradient Descent). MWE approach gave the best results in Task A, while, transformers and linguistic approach proved to be the best in Task B subtask 1 and 2 respectively.

[11] https://huggingface.co/docs

4. **Hate-alert:** Team Hate-alert participated in both tasks and worked with many transformer methods, however, the most successful methods proved to be mBERT (multilingual BERT) and MuRIL (Multilingual Representations for Indian Languages). For task A and task B subtask 2, mBERT produced the best results. On the other hand, MuRIL achieved the best results in Task B subtask 1.
5. **MUCS:** MUCS participated in both tasks and proposed a transfer learning model with mDistilBERT (Multilingual Distilled version BERT) and Classifier-chain model with SVM (linear Support Vector Classifier). The classifier-chain model is used for multi-label classification and used word n-grams (1-3) features to achieve the 4th rank in Task A. A fine-tuned mDistilBERT with LSTM as the final layer was used to achieve the best results for Task B.
6. **Aces:** The participants in team Aces experimented with four models for Task A including Recurrent Neural Network (RNN), Classifier Chains, Long Short-Term Memory (LSTM) and Multi-label K-nearest neighbours. For the deep learning models, they used fastText and Word2Vec, while for the machine learning models they used simple TF-IDF features. The best performing model on the test set was Classifier Chains with TF-IDF features giving the macro-F1 of 0.381. This approach stood last in the competition for multi-label emotion classification task.
7. **SakshiEmo2022:** Team Sakshi participated in both tasks of the challenge. The team ranked first in Task B subtask one and third in subtask two by using multiple fine-tuned transformer models. To handle the imbalanced classes in the task, the team used over-sampling and stratified sampling. The authors presented the results with MuRIL, BERT-base, Multilingual-BERT (mBERT), Distil-BERT and UrduHack. It was seen that MuRIL outperformed other transformers using the over-sampling method, whereas mBERT remained the next best method. In task A, team Sakshi experimented with transformers (UrduHack and mBERT) and machine learning models (Naive-Bayes, LinearSVC, Logistic Regression). The best results were achieved with LinearSVC using TF-IDF features and data decomposition (One vs Rest). This method ranked 7th in the official rankings.
8. **Ttdmnx:** The authors utilised pre-trained BERT models available on HuggingFace and achieved 4th and 5th positions in subtasks 1 and 2 of Task B. They simplified subtask 2 into a binary classification problem after separating the non-threatening tweets from the threatening tweets and applied the model to detect group vs individual threat tweets.

7. Results and Discussion

Table 3, 4 and 5 present the best results of the submitted systems in each task. The results of the participating team showed a significant increase in the F1 scores compared to the baseline values of Task A. The best performing team achieved 0.687 macro-F1 with four out of eight teams having more than 0.6 macro-F1 score. Weighted F1 scores were more competitive with the top 7 teams getting more than 0.6. The best-performing teams used transformer models which was a trend most of the teams carried with slight variations.

Task B comparatively had higher scores in the binary classification challenge, however, we saw that all teams score less than 0.6 F1 in the multi-class setup. In Task B, it was noticed

that models failed to differentiate “hate” from “threat” frequently. Examples of such cases are highlighted in the paper which was the third-best model presented in the shared task. The paper also highlighted confusion in task A between the labels “sadness” and “surprise” which was a frequent occurrence.

Similarly, team ERTIM documented ambiguities among the classes “anger” and “disgust” in their best model. they divided the errors in Task B into three different categories: lexical, phrasal and deictic. The authors noticed that the subjunctive mode for jussive phrases, second-person address, and future tense to express consequence was more employed in threat tweets. They observed the connection of abusive language with both classes of threat and were not able to identify MWE categories and phrase expressions of “non-threat” class. Team hate-alert made similar observations and reported that for threatening tweet detection, sometimes the presence of words such as “killing” made confusion.

Table 3
Each teams best run score in Task A

Rank	Team	Accuracy	Weighted-F1	Micro-F1	Macro-F1	Hamming loss
1	FOSUNlpTeam	0.636	0.759	0.759	0.687	0.088
2	UMUTeam	0.616	0.743	0.749	0.669	0.088
3	hate-alert	0.612	0.709	0.724	0.615	0.092
4	MUCS	0.582	0.696	0.692	0.603	0.113
5	ERTIM	0.593	0.699	0.72	0.599	0.0918
6	SakshiEmo2022	0.385	0.611	0.477	0.466	0.34
7	Aces	0.426	0.381	0.458	0.24	0.169

Table 4
Each teams best run scores in Task B subtask 1

Rank	Team	Macro-F1	Accuracy	ROC-AUC
1	SakshiEmo2022	0.716	0.738	0.729
2	hate-alert	0.716	0.737	0.729
3	ERTIM	0.689	0.723	0.690
4	ttdmnx	0.681	0.722	0.679
5	MUCS	0.626	0.641	0.648
6	Discovery	0.592	0.659	0.590
7	PYIP	0.436	0.644	0.494

8. Conclusion

Emotion and targeted abuse detection-based NLP tasks are overlooked frequently, especially in low-resource languages. This overview paper presents the findings of the shared task "EmoThreat: Emotions and Threat detection in Urdu". In this shared task, thirty-five different teams registered and ten teams submitted their proposed systems (runs). These teams used various techniques ranging from feature engineering to ML and DL algorithms. The approaches used include ensemble methods, deep learning methods and transformers. Mostly, teams used non-Urdu

Table 5
Each teams best run scores in Task B subtask 2

Rank	Team	Macro-F1	Accuracy	ROC-AUC
1	ERTIM	0.539	0.693	0.655
2	hate-alert	0.535	0.696	0.660
3	SakshiEmo2022	0.518	0.673	0.658
4	MUCS	0.419	0.618	0.566
5	ttdmnx	0.410	0.666	0.588
6	Discovery	0.374	0.630	0.552
7	PYIP	0.287	0.634	0.496

specialized transformers such as BERT and RoBERTa as well as Urdu-specialized transformers such as MuRIL, RoBERTa-urdu-small to achieve better results. FOSUNlpTeam outperformed all the proposed systems by using XLM-RoBERTa for Task A. Team SakshiEmo2022 and ERTIM produced the best results for Task B subtasks 1 and 2 respectively. These tasks aim to attract and encourage researchers working in different NLP domains to address multi-label emotion detection and threatening language detection problem in Urdu. It also helps to mitigate the abusive and threatening content on social media platforms.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

[1] F. Adeeba, S. Hussain, Experiences in building urdu wordnet, in: Proceedings of the 9th workshop on Asian language resources, 2011, pp. 31–35.
[2] D. Eberhard, G. Simons, C. Fennig, What are the Top 200 Most Spoken Languages, Ethnologue: Languages of the world (2021).
[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] S. Butt, F. Balouchzahi, G. Sidorov, A. Gelbukh, CIC@ PAN: Simplifying Irony Profiling using Twitter Data, in: CEUR Workshop Proceedings, volume 3180, CEUR-WS, 2022, pp. 2402–2410.
[5] 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.
[6] S. Butt, S. Sharma, R. Sharma, G. Sidorov, A. Gelbukh, What goes on inside Rumour and non-rumour Tweets and Their Reactions: A Psycholinguistic Analyses, Computers in Human Behavior (2022) 107345.
[7] 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.
[8] 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.
[9] S. Butt, N. Ashraf, G. Sidorov, A. F. Gelbukh, Sexism Identification using BERT and Data Augmentation-EXIST2021, in: IberLEF@ SEPLN, 2021, pp. 381–389.
[10] A. M. Kaplan, M. Haenlein, Users of the World, Unite! The Challenges and Opportunities of Social Media, Business horizons 53 (2010) 59–68.
[11] P. Ekman, Basic Emotions, Handbook of cognition and emotion 98 (1999) 16.
[12] 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.
[13] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, A. Gelbukh, Urdu Sentiment Analysis with Deep Learning Methods, IEEE Access 9 (2021) 97803–97812.
[14] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, A. Gelbukh, Multi-label Emotion Classification in Texts using Transfer Learning, Expert Systems with Applications 213 (2023) 118534.
[15] R. Plutchik, The Emotions, University Press of America, 1991.
[16] A. Chaturvedi, R. Sharma, minoffense: Inter-agreement hate terms for stable rules, concepts, transitivities, and lattices, in: IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2022.
[17] V. Bansal, M. Tyagi, R. Sharma, V. Gupta, Q. X. Xin, A transformer based approach for abuse detection in code mixed indic languages., in: Transactions on Asian and Low-Resource Language Information Processing., 2023. doi:https://doi.org/10.1145/3571818.
[18] 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.
[19] I. Ameer, G. Sidorov, H. Gomez-Adorno, R. M. A. Nawab, Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods, IEEE Access 10 (2022) 8779–8789.
[20] 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.