

# Learning Models for Emotion Analysis and Threatening Language Detection in Urdu Tweets

Asha Hegde, Hosahalli Lakshmaiah Shashirekha

*Department of Computer Science, Mangalore University, Mangalore, India*

## Abstract

The aim of Emotion Analysis (EA) task is to analyze and categorize the input text according to predefined sets of emotions. Recently, people have turned to social media to express their feelings, opinions, emotions about news, movies, products, services and so on. People's emotions may help governments, businesses, film producers, and others to devise strategies and make decisions for various activities. Threatening content identification aims to detect, abusive, offensive, and aggressive content in any text. With the growth of social media specifically Twitter, content for EA and detection of threatening text on Twitter is also increasing creating demand for tools that can analyze them efficiently. However, these tasks are challenging due to the complex nature of tweets. To tackle these issues, in this paper, we - team MUCS, describe two distinct models: i) Classifier-chain - a multi-label classifier using Support Vector Machine (SVM) and ii) Transfer Learning (TL) based model using Multilingual Distilled version Bidirectional Encoder Representations from Transformers (mDistilBERT) submitted to "EmoThreat: Emotions and Threat Detection in Urdu" shared task at Forum for Information Retrieval Evaluation (FIRE) 2022. Our models submitted to the shared tasks exhibited considerable results with an F1 score of 0.603 obtaining 4<sup>th</sup> rank in Task A and F1 scores of 0.626, and 0.307 obtaining 5<sup>th</sup> and 6<sup>th</sup> ranks in Subtask 1 and Subtask 2 respectively of Task B.

## Keywords

Urdu, Multi-label Classification, Multi-class Classification, Threatening Content, Emotional Analysis, Machine Learning, Transfer Learning

## 1. Introduction

In the internet era, social media platforms, such as Twitter, YouTube, Facebook, etc., are becoming the primary means of expressing emotions, opinions, and reviews about movies, products, etc. Emotions are psychological states that impact people and are frequently depicted in comments or reviews about movies, news, products, and so on. The comments/reviews include words having meanings, such as happiness, rage, joy, contempt, boredom, depression, etc. EA is the automatic analysis and classification of input text into one of the predefined sets of emotions, such as happy, sad, angry, fear, and so on. Analyzing text for emotions helps to predict market trends, capture the response of audience to movies, video songs, skits, news, and reality shows,

---


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

✉ [hegdekasha@gmail.com](mailto:hegdekasha@gmail.com) (A. Hegde); [hlsrekha@gmail.com](mailto:hlsrekha@gmail.com) (H.L. Shashirekha)

🌐 <https://mangaloreuniversity.ac.in/dr-h-l-shashirekha> (H.L. Shashirekha)



© 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)

identify key emotional triggers that change the users' mood, train chatbots, provide adaptive services based on the mood of customer/user, and so on [1].

In addition to constructive comments and reviews, the anonymity of social media users has enabled people to share objectionable content, such as hate speech, abusive and offensive comments, and threatening content targeting a group or an individual and spreading violence on social media platforms [2] [3]. Hence, it is necessary to identify such threatening content and exclude them from social media.

Most of the research works for EA and identification of threatening content in social media text focus on English, leaving the task aside for several other Indian languages and Urdu. In recent years, there has been an increase in the EA and identification of threatening text in Urdu due to the availability of a large volume of user-generated social media text. Twitter has emerged as a significant social media platform, with more than 353 million active users<sup>1</sup> from various ethnic, cultural, linguistic, and religious backgrounds to express their opinions [4]. Twitter allows to share short texts with a maximum length of 280 characters [2]. It was the first to use hashtags (#) - a short form of phrases or words preceded by hash signs, to highlight the importance of a specific topic. In addition to hashtags, tweets may also contain different slangs (ex: IDK → I don't know), words with recurrent characters (ex: gooooooood n8 → good night), abbreviations (ex: RT → Retweet), and re-tweets (reply to tweets) making the tweets more challenging to handle. An increasing number of Urdu users on Twitter and the growing number of posts and comments they share make it nearly impossible to manually track and control the content. Therefore, these comments should be analyzed automatically and filtered out. In addition, EA and identification of threatening content are open-ended issues because of the creative users' creative posts on Twitter.

To address the challenges of EA and identifying threatening content in Urdu tweets, in this paper, we - team MUCS, describe the models submitted to "EmoThreat: Emotions and Threat Detection in Urdu" at FIRE 2022. This task has two subtasks: i) Task A - a multi-label EA in Urdu and ii) Task B - has two subtasks: iia) Subtask 1 - a binary text classification to classify the given tweet as 'threatening' or 'non-threatening' and iib) Subtask 2 - a multi-class classification task to classify the given tweet into 'non-threatening'/'individual'/'group'. Two models: i) Classifier-chain multi-label classifier with SVM considering the combination of word n-grams in the range  $n = (1, 3)$  and Urdu word vectors as features and ii) TL based model with mDistilBERT are proposed to address Task A and Task B respectively. The sample Urdu tweets and their labels along with the English translations are given in Table 1.

The rest of the paper is structured as follows: Section 2 contains related works and Section 3 explains the methodology. Section 4 describes the experiments and outcomes, and the paper concludes in Section 5 with future work.

## 2. Related work

Several researchers have explored analyzing tweets written in English for various tasks, such as EA, Sentiment Analysis (SA), Fake news detection, etc. However, very few research works have addressed tweets written in Urdu for such tasks and few of the relevant works are described

---

<sup>1</sup><https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>

Emotion Analysis		
Comments in Urdu	Label	English translation
میں کبھی غصہ نہیں ہوا	Neutral	I have never been angry
شہزاد بھائی دہری مبارک ہوین پچھنا نا دہری کس چیز کی ہے 🤔 سچ میں مجھے دہری خوشی ہوئی ہے	Happiness	Shehzad bhai..double congratulations.. why double...honestly I have got double happiness
سر بات تکلیف کی نہیں حیرت کی ہے	Surprise	Sir this not something to be hurt about..it is surprising
م کو خوشی ملی بھی تو بس عارضی ملی لیکن جو غم ملے وہ غم جاوداں ملے	Sadness	Whatever happiness we got was transitory..but the sorrows were eternal
چودہ دن خوشی کا قرنطین	Surprise, Neutral	14 days quarantine for happiness
جسے پایا ہی نہیں اسے کھونے کا ڈر ہے	Fear	Why worry about losing someone who was never yours

Detection of Threatening and Non-threatening content		
Comments in Urdu	Label	English translation
پکواس مت کرو	Threatening, Individual	Don't talk nonsense
بھارت کی ایٹمی حملے کی دھمکی	Threatening, Group	Atomic bomb threat for India
پاک انگلینڈ سیریز کا شیڈول جاری ہو گیا	Non-threatening	Pak England series schedule has been published

**Table 1**

Sample text from the given dataset for EA and threatening content detection

below:

Ameer et al. [5] created multi-label code-mixed (English and Roman Urdu) Urdu dataset for EA. The dataset contains 11,914 code-mixed Urdu SMS messages collected from SMS-AP-18 corpus [6] and each SMS in the dataset is annotated into one of twelve emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral by a minimum of three annotators. They implemented Machine Learning (ML) algorithms (One Verses Rest (OVR) multi-label classifier with Naive Bayes (NB) and OVR with Support Vector Classifier (SVC)) and Deep Learning (DL) algorithms (Conventional Neural network (CNN), Recurrent Neural Network (RNN), Bidirectional RNN (BiRNN), Gated Recurrent Unit (GRU), Bidirectional GRU (BiGRU), and Long Short Term Memory (LSTM)) for EA. A combination of word tri-grams and character n-grams (n = 8) and Keras embeddings are used as features to train ML and DL models respectively. Further, they also implemented TL based models: Bidirectional Encoder Representations from Transformers (BERT) and Generalized Auto-Regressive model for Natural Language Understanding (XLNet) for EA. Among all the models, OVR with SVC obtained the highest micro F1 score of 0.67. Ashraf et al. [7] created a multi-label Urdu corpus for EA that consists of 6,043 tweets where each tweet is annotated into one of six emotions: anger, happiness, disgust, sadness, surprise, and fear. Each tweet is annotated by a minimum of 3 annotators and inter-annotators' agreement is computed using Cohan's Kappa coefficient resulting in 71% agreement. The authors implemented ML models (Random Forest (RF), Decision Tree (DT), Sequential Minimal Optimization (SMO), and Adaboost) and DL models (CNN, LSTM, and LSTM+CNN) for EA. They used stylometric-based features, namely: word count and word density, pre-trained word embeddings, and n-grams in the range n = (1, 4) and n = (3, 9) for

word and character n-grams respectively to train ML models and fastText word embeddings to train DL models. The authors also implemented a TL based classifier with Multilingual BERT (mBERT). RF classifier with word uni-grams outperformed the other models obtaining a micro F1 score of 0.60.

A multi-class Urdu corpus for Sentiment Analysis (SA) was created by Khan et al. [8] collecting 9,312 Urdu reviews from various websites belonging to different domains, such as food and beverages, movies and plays, software and apps, politics, and sports. These reviews were annotated into one of the three classes, namely: positive, negative, and neutral by 3 annotators. The annotated Urdu corpus was then used to implement ML models (SVM, NB, Adaptive Boosting (Adaboost), Multi Layer Perceptron (MLP), Logistic Regression (LR), and RF) and DL models (CNN, LSTM, BiLSTM, GRU and BiGRU) to perform SA. ML models are trained with word uni-grams, bi-grams, and tri-grams, and fastText word embeddings and DL models are trained using Keras embeddings and fastText word embeddings. Further, they also implemented a TL based classifier with mBERT, and among all the models, mBERT outperformed with an F1-score of 0.81. Hegde and Shashirekha [9] proposed an ensemble of RF, MLP, Gradient Boosting (GB), and Adaboost classifiers to classify the given Urdu text into 'fake' or 'real'. Using the combination of word uni-grams, character n-grams in the range  $n = (1, 3)$ , and fastText word vectors to train the ensemble classifier, they obtained a macro F1 score of 0.55 securing the 12<sup>th</sup> rank in the shared task.

An annotated corpus for threatening language detection and target identification in Urdu tweets was created by Amjad et al. [2]. Their corpus contains 3,564 Urdu tweets and each tweet was annotated by a minimum of 3 annotators. These Urdu tweets are classified into 'threatening' or 'non-threatening' classes and the threatening tweets are further categorized into 'group' or 'individual' classes. They implemented ML models (SVM, LR, RF, MLP, and Adaboost) and DL models (CNN and LSTM) for threatening language detection and target identification. ML models are trained with word n-grams in the range  $n = (1, 3)$ , character n-grams in the range  $n = (3, 6)$ , and fastText word embeddings. Keras embeddings and fastText word embeddings are used as features to train CNN and LSTM models respectively. Among all the models, SVM classifier trained using fastText word embeddings obtained a maximum macro F1 score of 0.71.

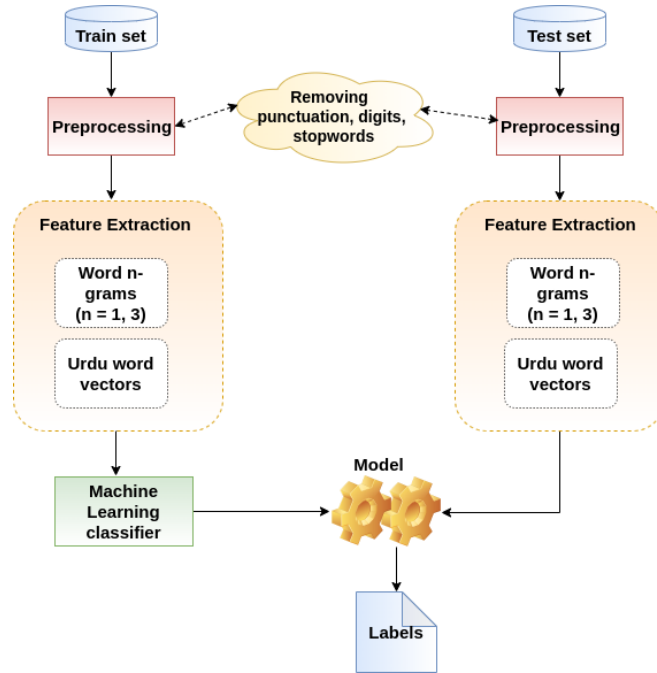
From the literature, it is clear that EA and the detection of threatening content in Urdu tweets are very less explored. Further, the models that are available for these tasks have shown less performance ensuring the scope for these tasks.

### 3. Methodology

The proposed methodology includes two distinct models, namely: Classifier-chain with SVM and TL model with mDistilBERT to address Task A and Task B respectively. Descriptions of the two models are given below:

#### 3.1. Classifier-chain with SVM model for Task A

The proposed Classifier-chain with SVM classifier consists of Pre-processing, Feature Extraction, and Model Building steps and the framework of the proposed model is shown in Figure 1. Each of the steps are explained below:



**Figure 1:** The proposed framework of ML classifier

**Preprocessing** - Punctuation, digits, URLs, and stopwords are removed from the dataset as they are ineffective for the classification task. Further, emojis are converted to English text and the Urdu stopwords list available at github<sup>2</sup> is used to remove the stopwords.

**Feature Extraction** - Word n-grams that express the relative importance between a word in the document and the entire corpus and pretrained Urdu word vectors<sup>3</sup> are used to train the linear SVC classifier for EA. Using word n-grams in the range  $n = (1, 3)$ , the total number of word n-grams extracted amounts to 30,410. The pretrained Urdu word vectors are trained with a window size 3, latent dimension 300 followed by a minimum count of 2.

**Model Building** - In addition to taking label dependencies into account for classification, Classifier-chain - a multi-label classifier, combines the computational efficiency of binary relevance methods [10]. This multi-label model arranges binary classifiers, such as SVM, LR, RF, etc., into a chain making the multi-label task easier. Each model makes a prediction in the order specified by the chain using all of the available features provided to the model plus the predictions of models that are earlier in the chain. SVM - an ML classifier, determines the decision boundary by the support vectors (optimal hyper-plane) to separate the output into different classes. In order to separate data points into classes, n-dimensional hyperplanes are drawn using the kernel trick. A kernel trick involves projecting nonlinear data onto a higher-dimensional space in order to make it easier to classify the data into areas where it can be linearly divided [11].

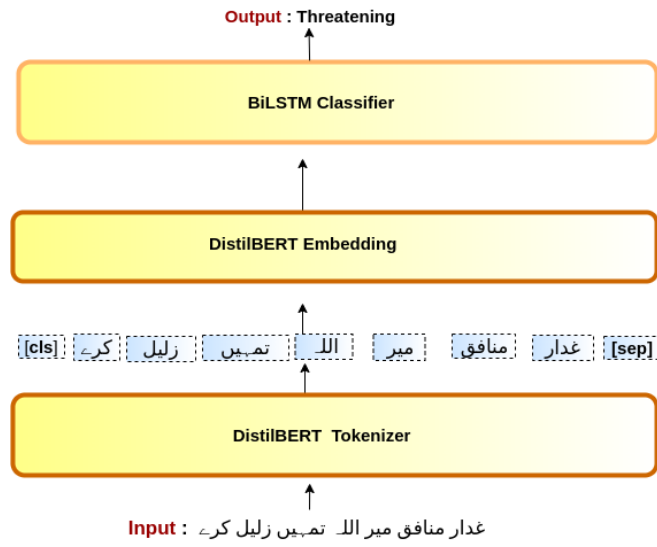
<sup>2</sup><https://github.com/stopwords-iso/stopwords-ur>

<sup>3</sup><https://github.com/samarh/urduvec>

mDistilBERT model	
Hyperparameters	Values
layers	6
dimension	768
heads	12
BiLSTM classifier	
layers	2
hidden units	256
dropout	0.03
learning rate	0.0001
activation function	relu

**Table 2**

Hyperparameters used in mDistilBERT based BiLSTM classifier and their values



**Figure 2:** The framework of TL model with BERT

### 3.2. Multilingual DistilBERT model for Task

The concept of TL is to train a model on one task and to make use of that model in a similar task. Instead of building the later model from scratch, the knowledge that is learned in one task is transferred to learn a similar task [12]. Multilingual DistilBERT - a TL-based model is implemented using the knowledge distillation of multilingual BERT (bert-base-multilingual-uncased) model that supports 100 languages, including Urdu [13]. To create a smaller version of BERT, mDistilBERT's creators removed the token-type embeddings and the pooler from the architecture and reduced the number of layers by a factor of 6. In this work, distilbert-base-multilingual-cased<sup>4</sup> model from the huggingface is used to extract the features. It may be noted

<sup>4</sup><https://huggingface.co/distilbert-base-multilingual-cased>

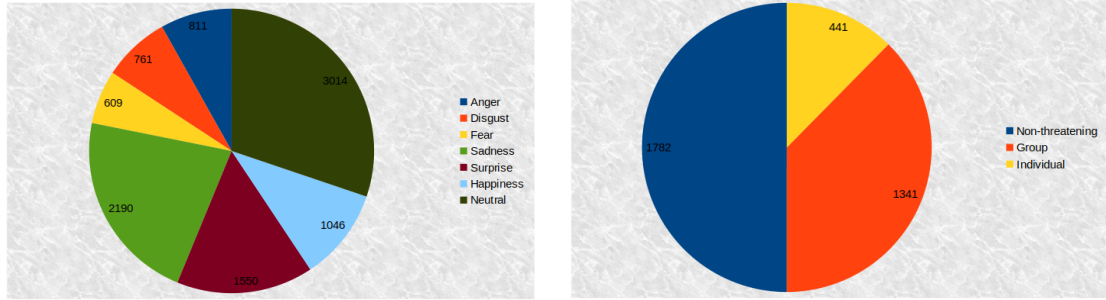
Task A	
Train set	7,800
Test set	1,950
Task B	
Train set	3,564
Test set	935

**Table 3**  
Statistics of the dataset

that the mDistilBERT is trained on a huge amount of unlabeled multilingual text available from various open-source corpora necessitating fine-tuning. Once the pretrained mDistilBERT model is loaded with the default parameter values, the model is frozen allowing the addition of a dense layer as the final layer to get the final output and then the model is retrained using the training data. It may be noted that, in this work, BiLSTM is used as the final layer for the prediction. Hyperparameters and their values used in mDistilBERT model are shown in Table 2 and the architecture is visualized in Figure 2.

## 4. Experiments and Results

The statistics of the dataset provided by the organizers of the shared task for EA<sup>5</sup> and identification of threatening content in Urdu text [2] for Task A and Task B are given in Table 3 and the classwise distribution of the datasets is given in Figure 3. Several experiments are conducted with various combinations of features and classifiers and the models that gave the good performance on the Development set are used to predict the labels of the Test set.



(a) Classwise distribution of Task A

(b) Classwise distribution of Task B

**Figure 3:** Classwise distribution of the dataset

The proposed models are used to predict the class labels of unlabeled Test sets provided by the organizers and the predictions were evaluated and ranked by the organizers based on the F1 score, separately for each task. Performance of the proposed models for Task A and B

<sup>5</sup><https://sites.google.com/view/multi-label-emotionsfire-task/dataset?authuser=0>

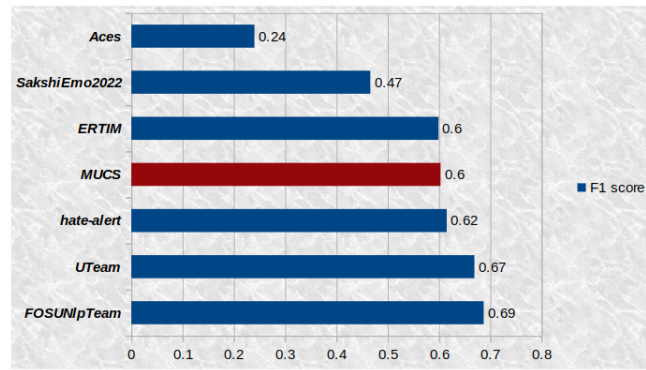


Task		F1 score	Rank
Task A		0.603	4
Task B	Subtask 1	0.626	5
	Subtask 2	0.307	6

**Table 4**  
Performance of the proposed models

along with the ranks obtained in the shared task are given in Table 4. The dataset for Task A and Subtask 2 of Task B are imbalanced whereas that of Subtask 1 of Task B is balanced. This imbalance may affect the performance of the classifier. The class imbalance problem of Task A is handled using the parameter 'class\_weight = balanced' in linear SVC.

The performance of both Task A and Task B are reported in Table 4. The proposed Classifier-chain model with SVM exhibited a considerable F1 score of 0.603 securing 4<sup>th</sup> rank in Task A. Further, the proposed mDistilBERT model exhibited F1 scores of 0.626 and 0.307 securing 5<sup>th</sup> and 6<sup>th</sup> rank for Subtask 1 and Subtask 2 respectively in Task B. Figures 4 and 5 illustrate the comparison of F1 scores of all the participating teams for Task A and B respectively demonstrating that the proposed models exhibited considerable performance.

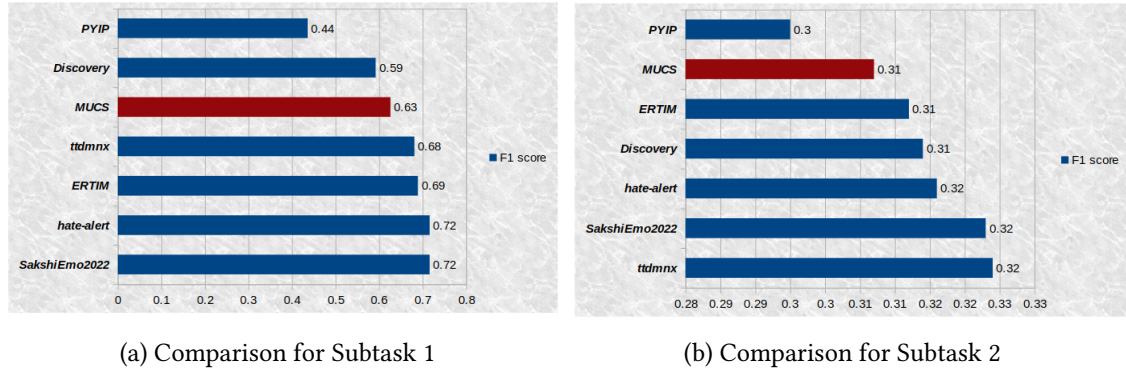


**Figure 4:** Comparison of F1 scores of the participating teams for Task A

## 5. Conclusion and Future work

This paper describes the models submitted by our team - MUCS to the shared task "EmoThreat: Emotions and Threat Detection in Urdu" at FIRE 2022 for EA and identification of threatening content in Urdu tweets. The two proposed models: i) Classifier-chain with SVM trained using the combination of word n-grams and Urdu word vectors and ii) TL model with mDistilBERT are proposed for Task A and B respectively. Classifier-chain model with SVM exhibited a considerable F1 score of 0.603 securing 4<sup>th</sup> rank in Task A and the proposed mDistilBERT model exhibited F1 scores of 0.626 and 0.307 securing 5<sup>th</sup> and 6<sup>th</sup> for Subtask 1 and Subtask 2 respectively in Task B. Future work will explore efficient resampling techniques for handling imbalanced classes with effective feature extraction.





**Figure 5:** Comparison of F1 scores of the participating teams for Task B

## References

- [1] A. Hegde, S. Coelho, H. Shashirekha, MUCS@DravidianLangTech@ACL2022: Ensemble of Logistic Regression Penalties to Identify Emotions in Tamil Text, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, 2022, pp. 145–150.
- [2] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening Language Detection and Target Identification in Urdu Tweets, in: IEEE Access, IEEE, 2021, pp. 128302–128313.
- [3] M. Das, S. Banerjee, P. Saha, Abusive and Threatening Language Detection in Urdu using Boosting based and BERT based Models: A Comparative Approach, in: arXiv preprint arXiv:2111.14830, 2021.
- [4] Y. Mehdad, J. Tetreault, Do Characters Abuse More Than Words?, in: Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue, 2016, pp. 299–303.
- [5] I. Ameer, G. Sidorov, H. Gomez-Adorno, R. M. A. Nawab, Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods, in: IEEE Access, IEEE, 2022, pp. 8779–8789.
- [6] M. Fatima, S. Anwar, A. Naveed, W. Arshad, R. M. A. Nawab, M. Iqbal, A. Masood, Multilingual SMS-based Author Profiling: Data and Methods, in: Natural Language Engineering, Cambridge University Press, 2018, pp. 695–724.
- [7] N. Ashraf, L. Khan, S. Butt, H.-T. Chang, G. Sidorov, A. Gelbukh, Multi-label Emotion Classification of Urdu Tweets, in: PeerJ Computer Science, PeerJ Inc., 2022, pp. 874–896.
- [8] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, Multi-class Sentiment Analysis of Urdu Text using Multilingual BERT, in: Scientific Reports, Nature Publishing Group, 2022, pp. 1–17.
- [9] A. Hegde, H. L. Shashirekha, Urdu Fake News Detection Using Ensemble of Machine Learning Models, in: CEUR Workshop Proceedings, 2021, pp. 132–141.
- [10] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier Chains for Multi-label Classification, in: Machine learning, Springer, 2011, pp. 333–359.
- [11] W. S. Noble, What is a Support Vector Machine?, in: Nature biotechnology, Nature Publishing Group, 2006, pp. 1565–1567.

- [12] B. Fazlourrahman, B. Aparna, H. Shashirekha, CoFFiTT-COVID-19 Fake News Detection Using Fine-Tuned Transfer Learning Approaches, in: Congress on Intelligent Systems, Springer, 2022, pp. 879–890.
- [13] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter, in: arXiv preprint arXiv:1910.01108, 2019.