Computers, Materials & Continua
DOI: 10.32604/cmc.2022.025543
Article

Tech Science Press

Deep Learning and Machine Learning-Based Model for Conversational Sentiment Classification

Sami Ullah[1], Muhammad Ramzan Talib[1,*], Toqir A. Rana[2,3], Muhammad Kashif Hanif[1] and Muhammad Awais[4]

[1]Department of Computer Science, Government College University Faisalabad, 38000, Pakistan
[2]Department of Computer Science and IT, The University of Lahore, 54590, Pakistan
[3]School of Computer Sciences, Universiti Sains Malaysia (USM), 11800, Penang, Malaysia
[4]Department of Software Engineering, Government College University Faisalabad, 38000, Pakistan
*Corresponding Author: Muhammad Ramzan Talib. Email: ramzan.talib@gcuf.edu.pk
Received: 27 November 2021; Accepted: 30 December 2021

Abstract: In the current era of the internet, people use online media for conversation, discussion, chatting, and other similar purposes. Analysis of such material where more than one person is involved has a spate challenge as compared to other text analysis tasks. There are several approaches to identify users' emotions from the conversational text for the English language, however regional or low resource languages have been neglected. The Urdu language is one of them and despite being used by millions of users across the globe, with the best of our knowledge there exists no work on dialogue analysis in the Urdu language. Therefore, in this paper, we have proposed a model which utilizes deep learning and machine learning approaches for the classification of users' emotions from the text. To accomplish this task, we have first created a dataset for the Urdu language with the help of existing English language datasets for dialogue analysis. After that, we have preprocessed the data and selected dialogues with common emotions. Once the dataset is prepared, we have used different deep learning and machine learning techniques for the classification of emotion. We have tuned the algorithms according to the Urdu language datasets. The experimental evaluation has shown encouraging results with 67% accuracy for the Urdu dialogue datasets, more than 10, 000 dialogues are classified into five emotions i.e., joy, fear, anger, sadness, and neutral. We believe that this is the first effort for emotion detection from the conversational text in the Urdu language domain.

Keywords: Dialogue analysis; conversational opinion mining; sentiment analysis; sentiment analysis in Urdu language; deep learning; machine learning

1 Introduction

Sentiment analysis (SA) has become an enduring field of research in the domain of text mining and Natural language processing (NLP). SA is the computational conduct of opinions, sentiments,

and partiality of the script [1–4]. Sentiment and topic analysis have been well discussed in the field of natural language processing [5]. The major goal of sentiment analysis is the identification of users' opinions, emotions, or sentiments from online text. These sentiments could be expressed for some organization, an individual, for some product or their aspects/attributes [6]. The massive growth of the internet and data availability over the web has opened new challenges for text analysis and automated systems.

SA comprises various fields elaborating the participation of emotions. These fields take account of Emotion detection (ED), Building resources (BR), and Transfer learning (TL). The major task of emotion analysis is to identify users' emotions which are hidden within the text written by the user against some entity. Transfer learning or Cross-Domain involves the learning the outcomes from one domain and then exploring a target domain according to the knowledge learnt from the first domain. Building Resources concerns with the development of the language or domain resource which may include generation of lexicons and corpus where expression are represented with respect their scope in particular domain, and occasionally vocabularies [7]. Emotion detection (ED) is considered as a specific division that lies under the sentiment analysis domain which can be described in a way to extract and analyze the emotions. ED evolved from texts and features the primary methodologies embraced by scientists in the plan of text-based ED systems. The proposition is examined comparable to their significant commitments, approaches utilized, datasets used, results acquired, qualities, and shortcomings [8].

Among different fields of SA, analyzing users' conversations has a great implication for commercial, academic, and government authorities. Due to the COVID-19 pandemic, the world has witnessed a new era of virtual lifestyle where everything from academia to healthcare, shopping to government activities, trading to conferences. There is a tremendous increase in the use of social media and online conversations have been witnessed. However, what is the impact of these online conversations or what kind of sentiments these conversations hold are the open challenges. Identification of such sentiments can be useful for identifying online bullies, criminal activities on social media, impact on the customer of online services, etc.

The Urdu language is considered as 21st language with the largest speakers of the world and also known as "Lashkari [لشکری]" language. An estimation of 66 million users are there around the world using Urdu language as a communication medium. However, the Urdu language is considered as resource-poor language as compared to the English language. This makes it more challenging to execute sentiment analysis task on the Urdu language corpora in the absence of standard language resources. In the past few years, we have noticed that obstinate postings over the web-based media have reshaped organizations, and influenced public feelings, which have significantly affected our social and political frameworks. Such postings have likewise activated masses for political changes, for example, those that occurred in some Arab nations in 2011. However, the Urdu language has not been explored much for the sentiment analysis especially there exists no work focusing on the dialogue analysis in the domain of Urdu language.

Building on the above explanation, this paper is designed to assimilate active dialogue learning that emerged from prior standard agenda in order to compare various common ensemble classifiers for dialogue sentiment classifications of reviews collected on multiple tasks. To the paramount of our understanding, there are most of the corpora published officially for the English language [9–13] however, we are not conscious of any reading that produced resources rather models to recognize sentiment stimulus in Urdu language using conversational datasets.

Therefore, in this paper, we have developed a dialogue dataset for the Urdu language and proposed a deep learning-based model for the sentiment classification of the developed Urdu dialogue dataset. As there exists no conversational dataset for the Urdu language, we acquired three benchmarked datasets and translated it into the Urdu language associated with the emotion labels. The text preprocessing has been performed for corresponding punctuation marks, hash tags, usernames, and every redundant junk including all the stop words. The processes of tokenization subsequently supplementary padding sequences were done appropriately then we obtained labels categorized whole of the processed, tokenized material as well as created embedding matrix. Moreover, we have utilized pre-trained language models to create embedding consequently amalgamated them with the dataset embedding. Afterward, we conceded the whole embedding to the deep learning models like CNN, LSTM, and BERT. We have evaluated the model on the developed datasets and have noticed the remarkable performance for predicting sentiments in the conversational datasets for the Urdu text.

The proposed model will help to explore the area of dialogue analysis for the users of the Urdu language. Due to the growing number of users over the internet and available data for regional languages, it has become essential to explore problems in such regional languages. Neglecting such languages results in a huge amount of data being unexplored. The dataset prepared can sever as a benchmark datasets for future research in this domain.

Based on the above, the following are the contributions of this research:

- We have developed the first conversational dataset for the Urdu language. The dataset is prepared by translating the three benchmarked datasets along with the emotion labeling. To the best of our knowledge, this is the first dataset of its kind in the Urdu language.
- We have presented a machine learning and deep learning-based model for the sentiment classification in the conversational dataset. The baseline results demonstrated the evaluation of state-of-the-art machine learning and deep learning algorithms in the Urdu language conversations dataset.

The rest of the paper is organized as follows: a comprehensive review of literature has been presented in Section 2. The proposed approach has been explained in Sections 3 and 4 presents the details of experiment conducted and obtained results and finally conclusion in Section 5.

2 Related Work

2.1 Dialogue Analysis

Deriu et al. [14] provided an overview of different datasets and matrices used for the evaluations for the dialogue analysis systems. Acheampong et al. [8] presented several open challenges and potential directions for text-based emotion detection. They aimed to present a guideline for the new researchers in the domain of text-based emotion detection and presented related theories, approaches, and labeled datasets available for executing text-based emotion detection tasks. Acheampong et al. [15] discussed latest articles where various BERT-based models have been proposed. They presented a discussion on the models using transformer-XL, cross-lingual language models, and the bidirectional encoder representations from transformers.

Hararika et al. [16] proposed a transfer learning-based model which pre-trained dialogue model and used conversational emotion classifier by transferring the parameters of the trained model. The proposed approach deduced information obtained by the conversation transformers which assist to predict the dialogue sentiments. Ghosal et al. [17] proposed a model based on the introduction of COSMIC, a judgment governed model in dialogue sentiment analysis. This model is based on the

immense commonsense information related to the character, special occasions, humor and mental processes, and expectations which are the cause of intense sentiments. Choi et al. [18] introduced a sentiment generation model, contributed emotional and natural responses as well by input utterances based upon the qualitative and quantitative analysis. They utilized a neural sentimental classifier contingent on Long short-term memory (LSTM) framework for the extraction of emotions from an emotion-labeled dialogue dataset.

An Adaptive Label Smoothing (AdaLabel) technique was proposed by Wang et al. [13] where the target label distribution was estimated adaptively. They used a bi-directional decoder facilitated by a novel target mask attention to generating context-sensitive control signals specifically towards no targeted words. Dang et al. [19] developed a dataset of German news headlines and labeled it with emotions and labeled instances of stimulus phrases. They used Conditional random field (CRF) for the classification of emotions and compared cross-lingual with their training model using projection. Demszky et al. [20] conducted experiments using transfer learning in existing emotion datasets to represent data generalization into different classifications and domains for example Tweets and personal stories. They conducted experiments for specific domains using additional emotion classification data; provided baseline emotions understanding to enhance model accuracy towards the targeted domain.

Oberländer et al. [9] prepared labeled dataset for English news headlines which was annotated using crowdsourcing with the related emotions, textual cues, associated with the causes of emotions, and their targets and also included the reader's perception. They designed a two-phase annotation mechanism for emotion constructions using crowdsourcing and provided the results of their baseline model for the prediction of such roles in a sequence labeling setting. A BERT-based evaluation metric called Dialog evaluation using BERT (DEB) was proposed by Sai et al. [21] and Xing et al. [22] which pre-trained Reddit conversations. Their model performed significantly better on random negatives and produced an accuracy of 88.27% in distinguishing the positive and random negative responses. A simple and effective strategy to generate a training corpus for utterance-pair coherence scoring was presented by [22]. They trained a BERT-based neural utterance-pair coherence model with the help of their trained corpus. Poria et al. [23] proposed the multimodal emotion-lines dataset (MELD) which include textual dialogues and also include visual and audio counterparts. MELD provided multimodal sources and was able to apply as a multimodal affective dialogue system for enhanced grounded learning. Batbaar et al. [24] proposed novel neural network architecture called Semantic-emotion neural network (SENN) which utilized pre-trained word representations for both the semantic/syntactic and emotional information. Pandelea et al. [25] proposed a novel framework which was based on the dual-encoder architecture for hardware-aware retrieval-based dialogue systems and grouped the candidates belonging to the same conversation using a clustering method.

2.2 Sentiment Analysis in the Urdu Language

There is no known work in the domain of dialogue analysis for the Urdu language because of the unavailability of language resources and datasets. There exists not a single dataset for dialogue analysis in the Urdu language and hence there is no work available. However, several approaches focusing on the sentiment analysis task in the Urdu language domain have been proposed. Rana et al. [26] proposed a rule-based model to identify opinions and their targets from the Urdu dataset. The rules used were manually crafted and a list of Urdu opinion words was utilized to identify opinion words. Amin et al. [27] used a ranking-based model to identify titles of the documents for the Urdu language. Rehmand et al. [28] used opinion lexicons for the classification of sentences and identification of opinions. Opinions were extracted with the help of opinion lexicons and each opinion was assigned a polarity score for

the classification of the sentences. Mukhtar et al. [29] also used opinion lexicons for the extraction of opinion terms and defined several rules to assign polarity scores to each extracted opinion.

Khan et al. [30] developed an Urdu language dataset for the sentiment analysis task. They presented a comparative analysis of the performance of different deep learning algorithms over Urdu datasets for sentiment analysis. Ali et al. [31] used word filtering and feature selection and optimization techniques to improve the detection of hate speech from Urdu Tweets. They used machine learning algorithms to classify the text. Awais et al. [32] exploited discourse information to improve the performance of the Urdu sentiment classifier. The Haruf-Ataf were used to identify opinions and their inner discourse relationships. Furthermore, they used a bag-of-words algorithm to identify the polarity relation between sub-opinions [33]. The overall polarity of the sentence was calculated with the help of a rule-based technique for the classification of sentiments with the help of calculated polarity, discourse relation, and polarity relation of two sub-opinions.

The above-mentioned approaches were applied to standard Urdu datasets however, several approaches have also focused on Roman Urdu sentiment analysis. Daud et al. [34] proposed RUOMIS, an opinion mining system for Roman Urdu. The reviews written in Roman Urdu script were extracted and translated into English language and then English language resources were used to calculate the sentiment orientation of each sentence. Nargis et al. [35] presented an ontology-based approach to classify Roman Urdu text according to the extracted emotions. They developed an emotion ontology that was capable to identify emotions using a sentiment analyzer, phrase analyzer, and format analyzer from Roman Urdu text. Mehmood et al. [36] developed a public dataset along with three pre-trained models for neural word embedding to accomplish the sentiment analysis task in Roman Urdu. They also evaluated different deep learning algorithms over the pre-trained embedding models.

Rana et al. [37] had proposed an unsupervised rule-based approach for opinion classification in the Roman Urdu dataset. They crafted rules manually and used several language techniques to correctly identify opinion words. Sohail et al. [38] used a photonics-based approach for the normalization of Roman Urdu words. They also utilized string similarity techniques along with the photonic algorithm. The collected dataset was divided into several classes and extracted sentiments were classified with the help pre-defined set of classification. Khan et al. [39] used a clustering-based technique to normalize Roman Urdu text and also used a photonic algorithm along with similarity metrics for the classification of the text.

The existing work in the Urdu language focused on the sentiment analysis task using review datasets. None of the existing work has focused on conversational datasets because no such dataset exists. Therefore, there is a need to build datasets which contain dialogues from the Urdu language. Hence, this study has focused on building such datasets and applying machine learning and deep learning approach to identify users' sentiments from these dialogues.

3 Methodology

This section highlights the flow of the proposed model for sentiment classification in the domain of the Urdu language. We have applied several preprocessing techniques to clean the data and to remove noises from the text. After preprocessing, we have applied tokenization and padding techniques to the input data. Once the dataset is prepared, deep learning algorithms have been utilized to classify the input data. Fig. 1 elaborates the sequence of the proposed model.

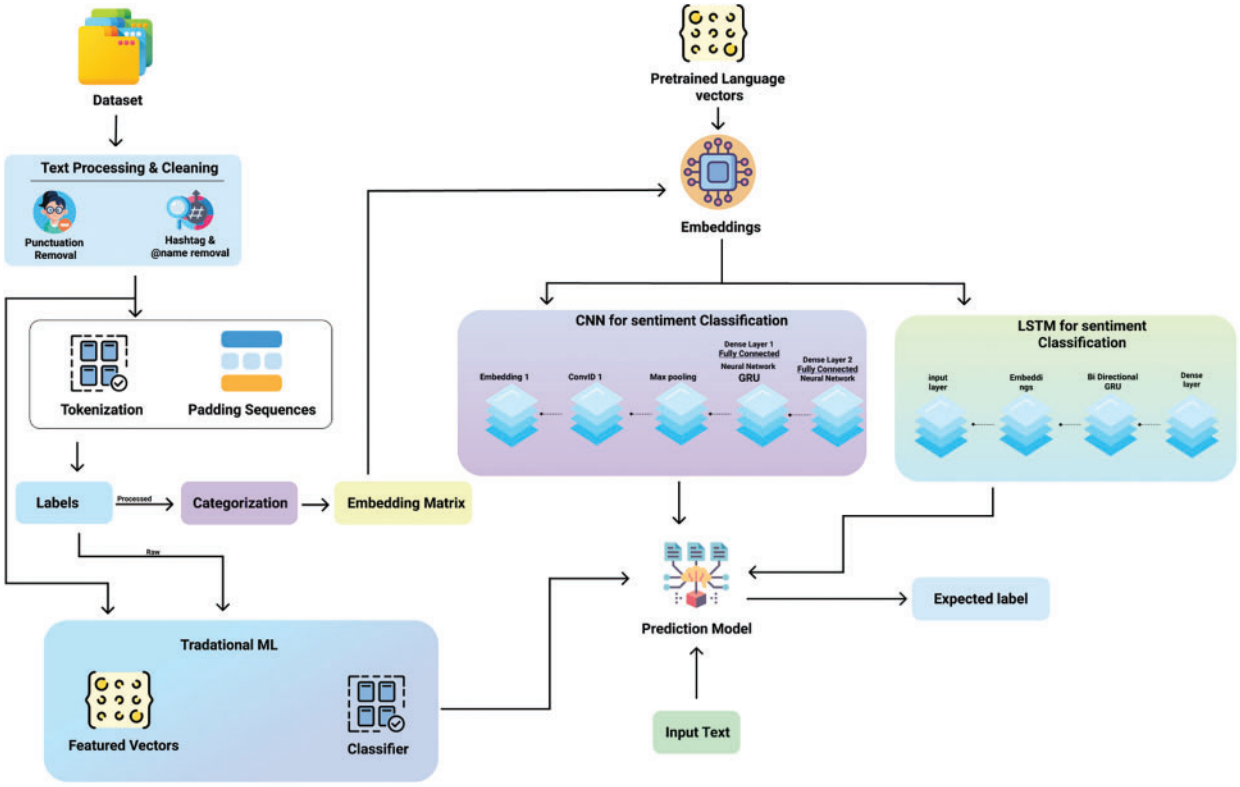

Figure 1: The proposed model

3.1 Corpus Generation

The Urdu language is considered a resource-poor language despite having a large number of users around the globe. This is due to the unavailability of language resources and the nonexistence of benchmark datasets. Although some datasets are available for sentiment analysis in the Urdu language [29] however there is no available dataset for dialogue analysis. One way is to build a new dataset which is a laborious task and there is not much material available on the internet for the Urdu language regarding conversations. The only possible way is to collect it manually and required human verification and annotation. Meanwhile, there are several datasets available for the English language which have been used by several researchers. DailyDialog dataset is one of them which contains a large number of dialogues tagged with emotions joy, anger, fear, neutral, and sadness. Therefore, we have used three datasets DailyDialog [40], Emotion-Stimulus [10], and ISEAR [41], and translated into the Urdu language using Google translator. The emotion tags are left untranslated as the real task is to identify emotions from the Urdu text and tagging does not matter whether in English language or Urdu. Fig. 2 shows an example of the dataset translated in the Urdu language while emotion tags are in the English language.

3.2 Preprocessing

The dataset contains punctuation marks, white spaces, special characters, etc. which affects the accuracy of the classification. Therefore, preprocessing includes the removal of such noisy terms and removing the stop words.

	Emotion	Text
0	neutral	...بہت سی دوسری پینٹنگز ہیں جو میرے خیال میں بہتر
1	sadness	... پھر بھی کتا بوڑھا اور کم قابل ہوچکا تھا ، اور
2	fear	...جب میں ٹکٹ کی ادائیگی کے بغیر ٹیوب یا ٹرین میں
3	fear	...یہ آخری کافی پریشانی کا باعث بن سکتا ہے اور شا
4	anger	...وہ ان میں سے کچھ کے ساتھ دکھاتی جانے والی قربت
5	sadness	...جب میرے اہل خانہ نے سنا کہ میری والدہ کے کزن ج
6	joy	...یہ جان کر کہ میں چینی اقاسیا کے لیے معیارات جم
7	anger	...ایک ترجمان نے کہا: 'گلین ناراض ہیں کہ ننے' انا
8	neutral	. جی ہاں
9	sadness	...جب میں جلتے لوگوں کو دیکھتا ہوں تو مجھے دکھ ہو
10	fear	...میں اپنی تحقیق کے لیے ہسپتال گیا تھا اور گھر پ
11	sadness	...ایک دن میں نے ایک دوست سے سنا کہ جس لڑکے سے می
12	joy	کونر کی آواز خوشگوار تھی۔
13	anger	...بہت مضحکہ خیز۔ آج تمہیں کیا ہوا ہے؟ آپ میرے سی
14	neutral	...شاید ہمیں مزید بچے پیدا کرنے کی ضرورت ہے! ٹینا

Figure 2: Explanation of the sample data

3.3 Tokenization and Padding

Tokenization is one of the key tasks while processing the raw text for sentiment analysis tasks. Tokenization is the task to separate text into tokens which are the smaller units of the input text. We have performed word-level tokenization and assigned tokens to each word in the document. The dataset contains sentences of different lengths, and this can affect the performance of the classification algorithms. Hence, we have applied the padding technique to normalize the length of each sentence. In this process, an additional zero is added for the sentence with short length and truncated the sentences which exceeded the given length.

3.4 Deep Learning Models

Deep learning has been applied as they perform high accuracy with low computation and proved be effective in text classification. The task is to identify terms related to users' sentiments and hence we have selected CNN as it performs better on such problems as compared to sequential modeling approaches lie Recurrent neural networks (RNN) [42]. We have used CNN separately and with a smaller part made on LSTM. The data loaded first go to the CNN model therefore, 2 dense layers develop the model more efficiently and it makes the processing in LSTM easier. Since a major part of the work is done by the dense layers in CNN, so only 1 dense layer is enough for the LSTM model.

A CNN and LSTM systems are included in our suggested model. The model receives the word embedding applied to flight reviews as input. As encoding words into vectors improves performance for NLP tasks, we employed the word embedding strategy with LSTM and CNN models. We utilize Keras to construct word embedding which turns the text data into vector values because dense layer

requires numbers as input. The embedding assumes that two words with comparable contexts have similar meanings and as a result produces vector representations.

The word vectors of vocabulary terms are first trained using the Word2Vec toolkit from a large corpus. The regional CNN model divides a given text into regions using a sentence as a region.

Once the word vectors have passed through a convolutional layer and a max-pooling layer in each region, meaningful emotional features can be retrieved. Local (regional) features are then successively combined across regions using LSTM to create a text vector for valence-arousal (VA) prediction. The explanation of each layer is in the upcoming sections.

3.4.1 Embedding Layer

We utilize Word2Vec in our framework to get the vectors for the words as input. Word2Vec is a neural network that analyses text before deep-learning algorithms handle it. The intention is to classify the sentences with CNN in this challenge however CNN cannot interpret the words as a human can. Word2Vec handles it by converting text into a vector that CNN can interpret. At the same time, the vector generated by Word2Vec can be used to indicate word distance i.e., when the meanings of two words are similar, their vectors' values are also comparable.

Word embedding is an exemplification of text where words have similar representation where they have similar meanings. We have used 300-dimensional word vectors pre-trained on Wikipedia articles. We can also train the Word2Vec model with our data however our dataset is quite small and trained word vectors might not be as good as using per-trained.

Through padding, a sentence or message is expanded into a fixed-length one specified by a few heuristic principles. The lengths of sentences fluctuate frequently in text classification. In this scenario, zero-padding results in a high amount of incorrect data reducing classifier performance. It also has a significant impact on the results of the LSTM and CNN family models in text classification, since it affects the pooling and weight update processes. For example, consider two statements with opposite sentiment polarities that each contain only one word and two words.

3.4.2 Max Pooling Layer

Nonlinear interactions can be accounted for in models that use an activation function. We used the ReLU-Rectified Linear Unit activation function after each Convolution process. ReLU has acquired a lot of traction in the previous few years. Because ReLU implements the function $y = \max(x, 0)$ and input and output size of this layer remains same. With this the decision function's nonlinear properties improved and also the entire network without disturbing the receptive fields of the convolution layer. Because the gradient is always high (equal to 1) and does not saturate when the neuron activates, it avoids and corrects the vanishing gradient problem.

With a 2×2 window size, we utilized the max-pooling function, which takes the largest element from the corrected feature map within that window.

3.4.3 Convolutional Layer 1D

1D convolutional neural networks (1D CNNs) is a modified form of 2D CNNs that was recently developed. In some applications with minimal labeled data and large signal fluctuations gathered from various sources, 1D CNNs were recently proposed and promptly attained state-of-the-art performance levels. The embedding layer turns the text's tokenized vector into a vector of the desired size. There is only one convolution layer in our CNN model. The input to this layer is convolved using pooling

layers, which decreases computation complexity and controls overfitting. The convolution filter has a size of 256; following the convolution layer, max pooling is done, and dropout is applied after the dense layers.

3.4.4 Fully Connected Layer

In a neural network, a dense layer transmits all of the preceding layer's outputs to all of its neurons, with each neuron delivering one output to each subsequent layer's neuron. The LSTM/GRU gives a deep representation of the fully connected layers and this representation transformed into the final output classes. Fully connected layers along with batch normalization and optionally dropout layers for regularization are the main components.

3.4.5 Output Layer

Depending on the input problem, the output layer can be comprised of either Sigmoid for binary classification or Softmax for both binary and multi-classification output. As we are dealing with the Classification problem, we have utilized a Softmax function to compute the probabilities for the output layers.

3.5 Machine Learning Models

We have used several machine learning algorithms for the classification which includes Naïve Bayes, random forest, linear regression and linear support vector. The main reason of using these algorithms is the size of the dataset. Machine learning algorithms perform better on smaller datasets and the idea is to analyze the performance of both machine learning and deep learning algorithms on the generated corpus for Urdu language.

4 Experimental Evaluation

This section presents the details of the experiments conducted on the Urdu language dataset with the proposed approach. We have highlighted the details of the datasets, used algorithms, and obtained results.

4.1 Corpus

To the best of our knowledge, there exists no dataset for the dialogue analysis in the Urdu language domain. Therefore, we have existing datasets for the English language and translated them into the Urdu language for our experiments. We have used three datasets DailyDialog [36], Emotion-Stimulus [37], and ISEAR [41] datasets, and translated them into Urdu as explained in Section 3.1. We have used their common emotions and neglected those which are common among these datasets. Tab. 1 presents the details of each dataset and its contents.

Table 1: Details of the datasets

Corpus attributes	Details used
Total number of datasets used	03
Dataset used	DailyDialog, emotion-stimulus, ISEAR
No. of Dialogs in DailyDialog	13,118
No. of Dialogs in Emotion-Stimulus	2414

(Continued)

Table 1: Continued

Corpus attributes	Details used
No. of dialogs in ISEAR	7666
Average sentences in all datasets	37333
Total number of Dialogs in mixed (for balancing) dataset	10,306

4.2 Evaluation Criteria

We have used common emotions of the three datasets explained in the previous section. There are five common emotions in the datasets which are joy, sadness, fear, anger, and neutral. Hence, we set our experimental settings according to these five emotions. We have combined dialogues with common emotions and used them as one dataset in our experiments. Training size is set to 80% and 20% of the dataset is used for testing. Tab. 2 highlights statistics used for the experimental evaluation.

Table 2: Experimental settings

Experimental parameters	Details used
Training dataset size	7932
Test dataset size	2374
Total number of emotions	5
Number of joy emotions	2125
Number of sadness emotions	2102
Number of fear emotions	1952
Number of anger emotions	2061
Number of neutral emotions	2066

For the performance evaluation of our model, we have used two performance metrics i.e., F1-score and accuracy to evaluate the goodness of the proposed approach. Accuracy is used to measure the effectiveness of the model in terms of patterns among variables in the datasets. While F1-score is used to measure the performance of the model by combining precision and recall. Following are the standard formulas for these metrics.

$$Accuracy\ (A) = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$F1\text{-}score = \frac{\mathrm{TP}}{TP\ + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})} \tag{2}$$

In the above formulas, TP is the true positive i.e., number of positive examples classified correctly, TN is the true negative i.e., number of negative examples classified correctly, FP is the false positive i.e., number of negative examples classified incorrectly and FN is the false negative i.e., number of positive examples classified incorrectly.

4.3 Classification Approaches

The details of the proposed design using deep learning and machine learning model have been highlighted in this section. To execute the deep learning models, we have customized the framework and presented the details of the experimental settings.

4.3.1 Deep Learning Approaches

We have used two deep learning algorithms CNN and LSTM for the classification of emotions from the Urdu dataset. Fig. 3 elaborates the proposed approach using CNN for classification.

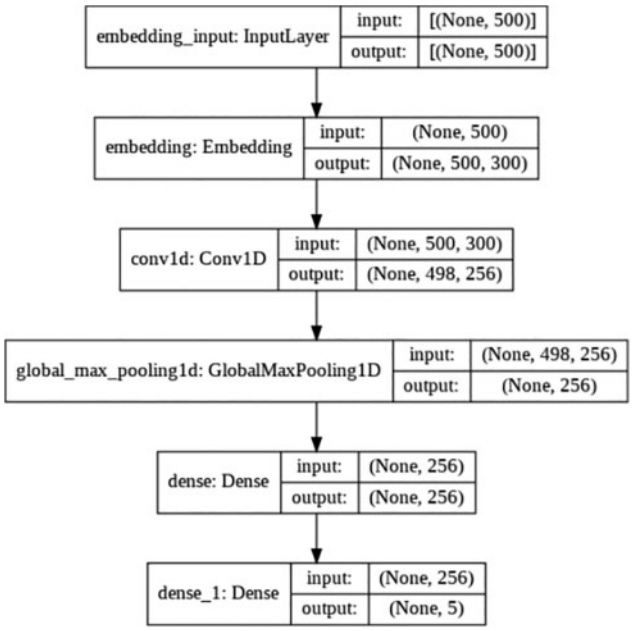

Figure 3: Process of CNN

The dataset after preprocessing is fed to the CNN where the number of dimensions for the embedding layer is set to 300. In our embedding layers, we have used the following pre-trained word vectors:

- Vocabulary size is 11509, which is the total number of words used for training and all the words are ignored after this limit.
- The maximum length in CNN model is 500 and this is the length of maximum input of the sequence.
- Embedding matrix in our case has the shape of (11509, 300), where 11509 is the number of entries (count or length) and 300 denote the dimensions.
- Train the model with 15 epochs.

The convolutional layer (CONV) performs convolutional operations by reading the input according to its dimensions and producing an output called activation map. The max-pooling layer selects the maximum value of the current view and transferred it to the fully connected layer where each input is connected to all neurons. On the other side, we have used bidirectional LSTM which uses input as sequence processing and uses both backward and forward sequences to increase the amount of information available. The process of the LSTM has been elaborated in Fig. 4.

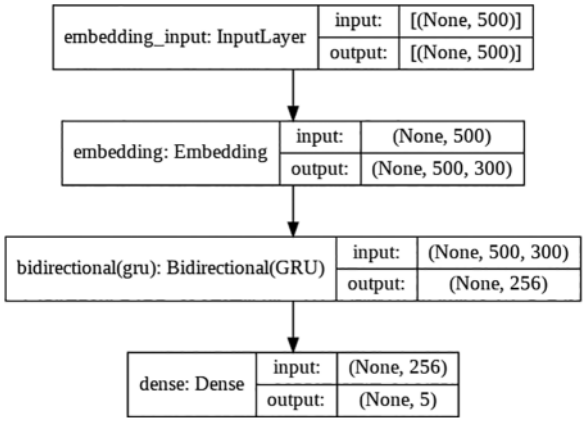

Figure 4: Flow of bidirectional LSTM

Fig. 5a presents the training accuracy of the deep learning approaches. The x-axis shows the number of epochs while the accuracy of the training is shown on the y-axis. It can be observed the training curve gradually increases and then stabilizes. It is clear from Fig. 5b that LSTM proved to be more efficient as compared to the CNN for dialogue analysis in the Urdu language.

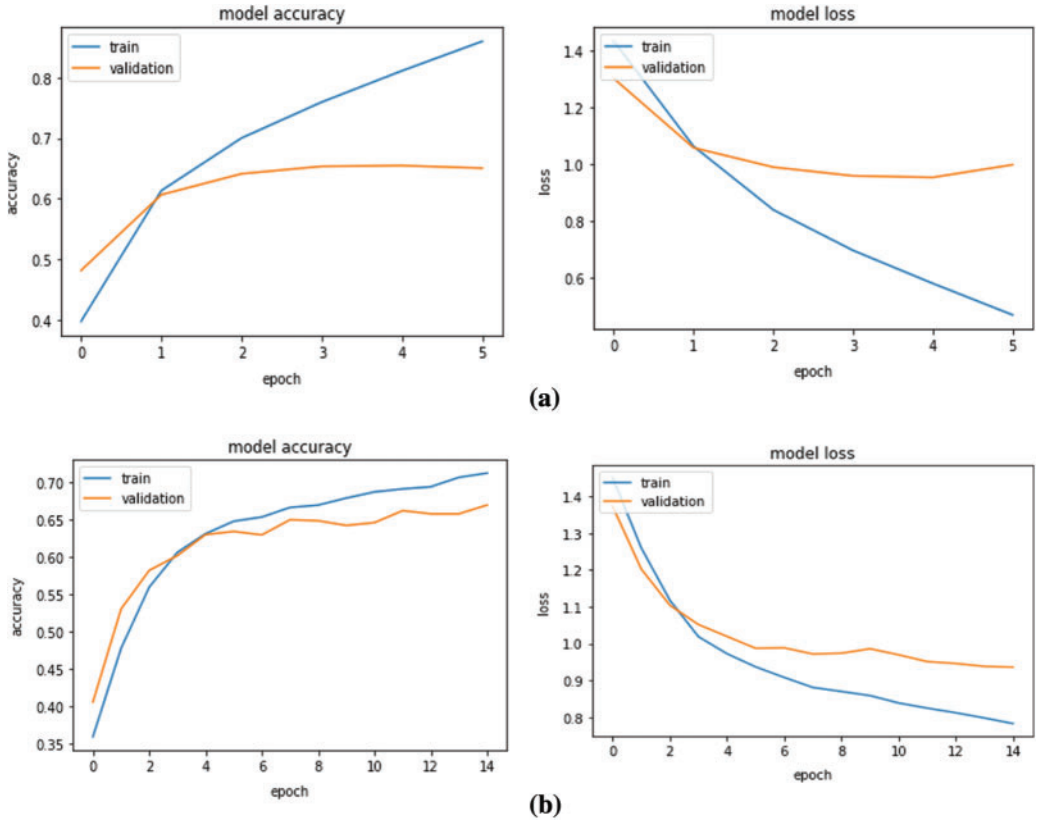

Figure 5: (a) Training accuracy of the CNN model (b) Training accuracy of the CNN model

4.3.2 Machine Learning Approaches

We have used four classification techniques Naïve Bayes (NB), Linear regression (LR), Random forest (RF), and Linear support vector (LSR). We have used 70% of the dataset as training and 30% for testing purpose.

4.4 Results and Discussion

In this section, we have summarized the results of deep learning and machine learning algorithms. Tab. 3 illustrates the performance evaluation of machine learning algorithms for each emotion. The best performance is highlighted in bold for each emotion. This is clear from the table that LSR has shown the best results for all the emotions except for "joy" and "fear" where NB has achieved the best accuracy. Similarly, Tab. 4 illustrates the performance evaluation of deep learning algorithms for each emotion. CNN has shown similar results to LSTM however its shows poor accuracy for "neutral" and hence LSTM has shown better results.

Table 3: Classification report for the machine learning algorithms

Emotion	Naïve bayes		Linear regression		Random forest		Linear support vector	
	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score
Joy	**0.68**	0.66	0.65	0.69	0.59	0.61	0.67	**0.72**
Sadness	0.50	0.61	0.63	0.65	0.55	0.58	**0.66**	**0.66**
Anger	0.69	0.62	0.67	0.64	0.60	0.56	**0.70**	**0.65**
Neutral	0.57	0.50	**0.77**	0.61	0.76	0.58	**0.77**	**0.63**
Fear	**0.72**	0.59	0.59	0.65	0.55	0.55	0.63	**0.67**

Table 4: Classification report for the deep learning algorithms

Emotion	CNN		LSTM	
	Accuracy	F_1-score	Accuracy	F_1-score
Joy	**0.72**	0.67	0.60	**0.68**
Sadness	0.73	**0.65**	**0.74**	0.62
Anger	**0.75**	0.61	0.74	**0.62**
Neutral	0.47	0.67	**0.62**	**0.74**
Fear	0.58	**0.64**	**0.66**	0.59

Fig. 6 shows the confusion matrices of all the models. CNN has produced poor results for "neutral" emotion as compared to LSTM while linear regression and linear support vector have shown consistent results for all the emotions.

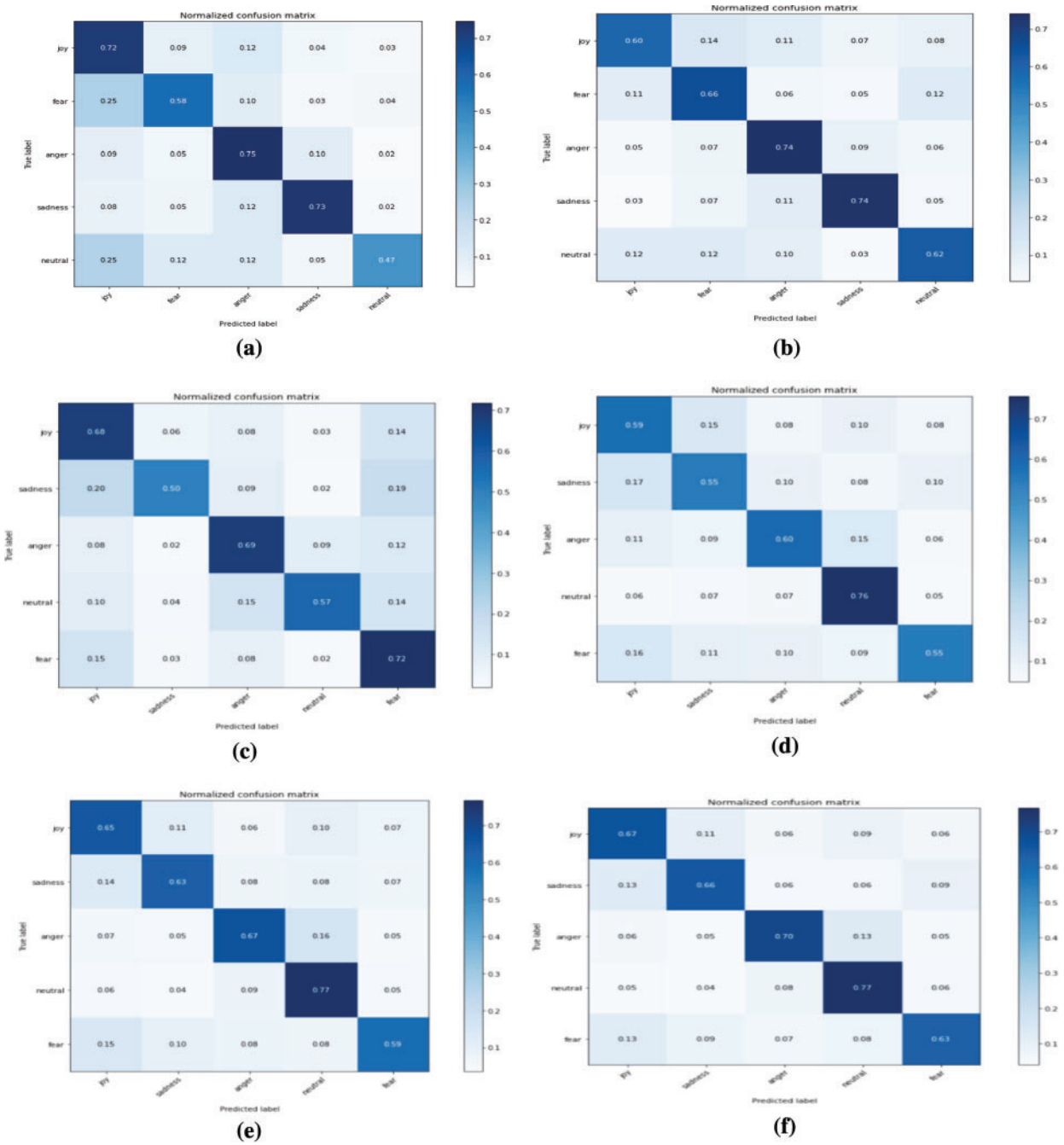

Figure 6: Confusion matrices for different classification algorithms (a) CNN (b) LSTM (c) Naïve bayes (d) Random forest (e) Linear regression (f) Linear support vector

Tab. 5 highlights the results of the different classification algorithms. This can be observed that LSTM has produced better results than CNN. However, LSV has shown the highest accuracy among all the classification algorithms. LSTM and LSV have comparable performance among all the classification techniques.

Table 5: Comparison of different classification algorithms

Algorithm	Accuracy	F_1-score
CNN	65.04	64.84
LSTM	67.02	67.05
Naïve bayes	59.63	57.58
Linear regression	65.40	65.43
Random forest	58.57	58.53
Linear support vector	67.45	67.49

5 Conclusion

The Urdu language is one of the most spoken languages in the South Asia region with an estimation of 66 million users around the world. With such a huge audience, there is not much work available for the text analysis especially with the best of our knowledge there is no dataset nor any work available for emotion identification from the conversational text in the Urdu language. Therefore, in this paper, we have generated the first dataset for dialogue analysis with the help of existing available datasets for the English language. The English datasets were already labeled with emotion and hence we have used the same labeling after translating the text into the Urdu language. After that, we have proposed a classification model which used deep learning and machine leering algorithms for the classification of emotions. Before applying classification, we have preprocessed that data and removed the noises and standardized the sentences. The results have shown that we have achieved encouraging results for the emotion classification task using a conversational dataset in the Urdu language domain. This work is the first effort for Urdu Language dialogue analysis. Overall, deep learning algorithms have produced consistent performance as compared to machine learning algorithms. The performance of deep learning algorithms can be further improved with the help of language models. However, there are no such language resources available for the Urdu language and future work may include building such language resources.

Acknowledgement: Author shall be thankful to Government College University for providing resources for this research.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] T. A. Rana and Y. N. Cheah, "Aspect extraction in sentiment analysis: Comparative analysis and survey," *Artificial Intelligence Review*, vol. 46, no. 4, pp. 459–483, 2016.

[2] T. A. Rana, Y. N. Cheah and S. Letchmunan, "Topic modeling in sentiment analysis: A systematic review," *Journal of ICT Research & Applications*, vol. 10, no. 1, pp. 76–93, 2016.

[3] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, Boston, MA: Springer, pp. 415–463, 2012.

[4] B. Liu, "Opinion mining and sentiment analysis," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–184, 2012.

[5] H. Saif, Y. He, M. Fernandez and H. Alani, "Contextual semantics for sentiment analysis of twitter," *Information Processing & Management*, vol. 52, no. 1, pp. 5–19, 2016.
[6] E. Fersini, "Sentiment analysis in social networks: A machine learning perspective," in *Sentiment Analysis in Social Networks*, Boston: Morgan Kaufmann, pp. 91–111, 2017.
[7] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath and A. Perera, "Opinion mining and sentiment analysis on a twitter data stream," in *Proc. of the Int. Conf. on Advances in ICT for Emerging Regions*, Colombo, Sri Lanka, pp. 182–188, 2012.
[8] F. A. Acheampong, C. Wenyu and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, pp. 1–24, 2020.
[9] L. A. M. Oberländer, E. Kim and R. Klinger, "Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception," in *Proc. of the 12th Language Resources and Evaluation Conf.*, Marseille, France, pp. 1554–1566, 2020.
[10] D. Ghazi, D. Inkpen and S. Szpakowicz, "Detecting emotion stimuli in emotion-bearing sentences," in *Proc. of the Int. Conf. on Intelligent Text Processing and Computational Linguistics*, Konya, Turkey, Springer, Cham, pp. 152–165, 2015.
[11] L. Gui, R. Xu, Q. Lu, D. Wu and Y. Zhou, "Emotion cause extraction, a challenging task with corpus construction," in *Proc. of the 5th Chinese National Conf. on Social Media Processing*, Nanchang, China, pp. 98–109, 2016.
[12] S. Poria, E. Cambria, L. W. Ku, C. Gui and A. Gelbukh, "A Rule-based approach to aspect extraction from product reviews," in *Proc. of the Second Workshop on Natural Language Processing for Social Media*, Dublin, Ireland, pp. 28–37, 2014.
[13] Y. Wang, Y. Zheng, Y. Jiang and M. Huang, "Diversifying dialog generation via adaptive label smoothing," in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing*, Bangkok, Thailand, vol. 2021, pp. 3507–3520, 2021.
[14] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset *et al.,* "Survey on evaluation methods for dialogue systems," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 755–810, 2021.
[15] F. A. Acheampong, H. N. Mensah and W. Chen, "Transformer models for text-based emotion detection: A review of bert-based approaches," *Artificial Intelligence Review*, vol. 54, pp. 5789–5829, 2021.
[16] D. Hazarika, S. Poria, R. Zimmermann and R. Mihalcea, "Conversational transfer learning for emotion recognition," *Information Fusion*, vol. 65, no. 1, pp. 1–12, 2021.
[17] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea and S. Poria, "COSMIC: Commonsense knowledge for emotion identification in conversations," in *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing: Findings*, Stroudsburg, Pennsylvania, United States, pp. 2470–2481, 2020.
[18] H. J. Choi and Y. J. Lee, "Deep learning based response generation using emotion feature extraction," in *Proc. of the 2020 IEEE Int. Conf. on Big Data and Smart Computing*, Busan, South Korea, pp. 255–262, 2020.
[19] B. M. D. Dang, L. Oberländer and R. Klinger, "Emotion stimulus detection in German news headlines," in *Proc. of the 17th Conf. on Natural Language Processing*, Düsseldorf, Germany, pp. 73–85, 2021.
[20] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade *et al.,* "Goemotions: A dataset of fine-grained emotions," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, Pennsylvania, United States, pp. 4040–4054, 2020.
[21] A. B. Sai, A. K. Mohankumar, S. Arora and M. M. Khapra, "Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining," *Transactions of the Association for Computational Linguistics*, vol. 8, no. 1, pp. 810–827, 2020.
[22] L. Xing and G. Carenini, "Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring," in *Proc. of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Singapore, pp. 167–177, 2021.
[23] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria *et al.,* "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, Pennsylvania, United States, pp. 527–536, 2018.

[24] E. Batbaatar, M. Li and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019.
[25] V. Pandelea, E. Ragusa, T. Young, P. Gastaldo and E. Cambria, "Toward hardware-aware deep-learning-based dialogue systems," *Neural Computing and Applications*, pp. 1–12, 2021. https://doi.org/10.1007/s00521-020-05530-1.
[26] T. A. Rana, B. Bakht, M. Afzal, N. A. Mian, M. W. Iqbal *et al.,* "Extraction of opinion target using syntactic rules in urdu text," *Intelligent Automation & Soft Computing*, vol. 29, no. 3, pp. 839–853, 2021.
[27] A. Amin, T. A. Rana, N. A. Mian, M. W. Iqbal, A. Khalid *et al.,* "Top-rank: A novel unsupervised approach for topic prediction using keyphrase extraction for urdu documents," *IEEE Access*, vol. 8, pp. 212675–212686, 2020.
[28] Z. U. Rehman and I. S. Bajwa, "Lexicon-based sentiment analysis for urdu language," in *Proc. of the Sixth Int. Conf. on Innovative Computing Technology*, Dublin, Ireland, pp. 497–501, 2016.
[29] N. Mukhtar, M. A. Khan and N. Chiragh, "Lexicon-based approach outperforms supervised machine learning approach for urdu sentiment analysis in multiple domains," *Telematics and Informatics*, vol. 35, no. 8, pp. 2173–2183, 2018.
[30] L. Khan, A. Amjad, N. Ashraf, H. T. Chang and A. Gelbukh, "Urdu sentiment analysis with deep learning methods," *IEEE Access*, vol. 9, pp. 97803–97812, 2021.
[31] M. Z. Ali, A. E. Haq, S. Rauf, K. Javed and S. Hussain, "Improving hate speech detection of urdu tweets using sentiment analysis," *IEEE Access*, vol. 9, pp. 84296–84305, 2021.
[32] D. M. Awais and D. M. Shoaib, "Role of discourse information in urdu sentiment classification: A rule-based method and machine-learning technique," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 18, no. 4, pp. 1–37, 2019.
[33] M. Hassan and M. Shoaib, "Opinion within opinion: Segmentation approach for urdu sentiment analysis," *International Arab Journal of Information Technology*, vol. 15, no. 1, pp. 21–28, 2018.
[34] M. Daud, R. Khan and A. Daud, "Roman urdu opinion mining system (RUOMiS)," *Computer Science and Engineering: An International Journal*, vol. 4, no. 6, pp. 1–9, 2014.
[35] G. Z. Nargis and N. Jamil, "Generating an emotion ontology for roman urdu text," *International Journal of Computational Linguistics Research*, vol. 7, no. 3, pp. 83–91, 2016.
[36] F. Mehmood, M. U. Ghani, M. A. Ibrahim, R. Shahzadi, W. Mahmood *et al.,* "A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis," *IEEE Access*, vol. 8, pp. 192740–192759, 2020.
[37] T. A. Rana, K. Shahzadi, T. Rana, A. Arshad and M. Tubishat, "An unsupervised approach for sentiment analysis on social media short text classification in roman urdu," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 2, pp. 1–16, 2021.
[38] O. Sohail, I. Elahi, A. Ijaz, A. Karim and F. Kamiran, "Text classification in an under-resourced language via lexical normalization and feature pooling," in *Proc. of the Twenty-Second Pacific Asia Conf. on Information Systems*, Japan, pp. 96, 2018.
[39] A. R. Khan, A. Karim, H. Sajjad, F. Kamiran and J. Xu, "A clustering framework for lexical normalization of roman urdu," *Natural Language Engineering*, vol. 28, no. 1, pp. 1–31, 2020.
[40] Y. Li, H. Su, X. Shen, W. Li, Z. Cao *et al.,* "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proc. of the Eighth Int. Joint Conf. on Natural Language Processing*, Florence, Italy, pp. 986–995, 2017.
[41] S. Poria, A. Gelbukh, E. Cambria, A. Hussain and G. B. Huang, "Emosenticspace: A novel framework for affective common-sense reasoning," *Knowledge-Based Systems*, vol. 69, no. 1, pp. 108–123, 2014.
[42] V. Gupta, N. Jain, S. Shubham, A. Madan, A. Chaudhary *et al.,* "Toward integrated CNN-based sentiment analysis of tweets for scarce-resource language—Hindi," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1–23, 2021.