ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 11 N. 4 (2022), 489-505
eISSN: 2255-2863
DOI: https://doi.org/10.14201/adcaij.30128

Contextual Urdu Text Emotion Detection Corpus and Experiments using Deep Learning Approaches

Muhammad Hamayon Khan Vardag[a], Ali Saeed[b],
Umer Hayat[a], Muhammad Farhat Ullah[a], Naveed Hussain[b]

[a]Department of Software Engineering, University of Lahore, Lahore, Pakistan
[b]Department of Software Engineering, Faculty of Information Technology, University of Central Punjab, Lahore, Pakistan
humayonverdag98@gmail.com, ali.saeed@ucp.edu.pk, hayatumer575@gmail.com, farhat.ullah@se.uol.edu.pk, and dr.naveedhussain@ucp.edu.pk

KEYWORDS

emotion detection corpus; labeled corpus; deep learning approaches; supervised learning approaches

ABSTRACT

Textual emotion detection aims to discover human emotions from written text. Textual emotion detection is a significant challenge due to the unavailability of facial and voice expressions. Considerable research has been done to identify textual emotions in high-resource languages such as English, French, Chinese, and others. Despite having over 300 million speakers and large volumes of literature available online, Urdu has not been properly investigated for the textual emotion detection task. To address this gap, this study makes two contributions: (1) the creation of a novel dialog-based corpus for Urdu (Contextual Urdu Text Emotion Detection Corpus). CUTEC contains 30,160 training and 5,509 testing labelled dialogues, where each dialogue consists of three Urdu contextual sentences. In addition, all dialogues are labelled using four emotion classes, i.e., Happy, Sad, Angry, and Other. As a second contribution (2) five deep learning models, i.e., RNN, LSTM, Bi- LSTM, GRU, and Bi-GRU have been trained and tested using CUTEC with different parametric settings. The highest results (Accuracy = 87.28 and F_1 = 0.87) are attained using a GRU-based architecture.

1. Introduction

Textual emotion detection in computational linguistics is the process of recognizing discrete emotions indicated in a text (Chatterjee, et al. 2019). Researchers widely studied four to six emotions from the text, i.e., happy, sad, angry, love, fear, and others (Baali and Ghneim 2019) (Bashir, et al. 2022).

The “Others” class is assigned to those particular text dialogues which do not enclose any emotion. Recently, EmoContext used these four classes for the textual emotion detection task (Chatterjee, et al. 2019). Human readers are normally capable of extracting emotions from the written line or a dialogue; however, doing so is generally hard for machines.

Textual emotion detection has the potential to assist multiple fields of computer science and information technology, such as human-computer interaction (Kao, et al. 2009) (Rincon, et al. 2016), dialogue classification (Baali and Ghneim 2019) (Bashir, et al. 2022), comment classification (Nagwani 2015), e-learning review improvement (V. Chang 2016), brand review identification (Syed and others 2015), detection of emotions on political events (Chang and Masterson 2020), product refinements through user reviews (Druck and Pang 2012), monitoring hostile and criminal communications (Panko and Beh 2002).

Generally, emotions are expressed through face, voice, body language, and text. Detecting emotions from facial expressions (Rani and Garg 2014) and speech (Yu, et al. 2001) has been widely explored by researchers, and now researchers are focusing on the text (Chatterjee, et al. 2019) (Bashir, et al. 2022) (Vijay, et al. 2018). Context (previous sentences) changes the emotion perceived through a sentence (Chatterjee, et al. 2019). Context can also decrease the ambiguity among emotions contained by the text, e.g., “ مجھے رونا پسند ہے” often classified as Sad, but a dialogue User-1: “مجھے رونا پسند ہے” User- 2: “تم کیوں رو رہی ہو؟” User-1: “یہایک مذاق تھا” can be categorized as Happy. Therefore, it is clear from the above example of three Urdu sentences, that context has a significant impact on emotion detection and has the potential to change the class. The focus of this research work is on contextual emotion detection task for the Urdu language. Urdu is an important South Asian language with 300 million users (Hussain 2008) (Khan, et al. 2016) (Riaz 2010). There are millions of L1 (151 million) and L2 (149 million) speakers of Urdu (Rahman 2004). Urdu is Pakistan’s official national language with 11 million L1 speakers (Rahman 2004) (Naseer and Hussain 2009). Other countries with a large community of Urdu speakers include India, Bangladesh, the USA, Canada, and Europe (Hussain 2008) (Khan, et al. 2016). Urdu is considered morphologically rich, which means, many of its words can have more than 40 interpretations. Consequently, it is difficult to process as compared to morphologically poor languages (such as English) (Hussain 2008) (Naseer and Hussain 2009).

Benchmarked labeled corpora are required to construct, analyze and evaluate textual emotion detection systems. A range of corpora constructed for Textual Emotion Detection tasks is for English and other popular international languages (Chatterjee, et al. 2019) (Baali and Ghneim 2019) (Bashir, et al. 2022) (Syed and others 2015) (Syed, et al. 2010). Whereas, there is a need for standard benchmarked labeled corpora for Urdu (a poorly resourced language). This research work represents a novel benchmark corpus and designed various deep learning models for contextual Textual Emotion Detection task for Urdu.

We chose deep neural networks because of their ability to comprehend data's internal representation with multiple hidden layers that do not require handcrafted feature engineering (Baali and Ghneim 2019) (ayir, Yenidouan and Da 2018).

Our Contextual Urdu Text Emotion Detection Corpus (CUTEC) contains 35,509 labeled dialogues. Following the standard practice of EmoContext- SemEval-2019 Task 3, a single dialogue is composed of three contextual sentences. EmoContext corpus has been used as a source corpus and all instances were automatically translated using Google Translator. Furthermore, the quality of each dialogue has been ensured by three human language experts. Furthermore, five deep learning-based approaches have been designed and evaluated against CUTEC to clarify how our proposed dataset can be utilized for the construction and deployment of an automatic contextual emotion detection system for Urdu.

Overall, this research presents two contributions: (1) a novel Urdu dialog-based corpus (Urdu Text Emotion Detection Corpus). Every dialogue in CUTEC has three contextual Urdu sentences. There are 30,160 training dialogues and 5,509 testing dialogues. All dialogues are also labelled with one of four mood categories: happy, sad, angry, or other. (2) The second contribution is the training and testing of five deep learning models using CUTEC, including RNN, LSTM, Bi-LSTM, GRU, and Bi-GRU. The most accurate results were obtained using a GRU-based architecture (Accuracy = 87.28 and $F_1 = 0.87$).

The rest of this paper is organized as follows: Section 2 presents a review of existing contributions on the topic of emotion detection in various languages. Section 3 describes the generation process of the proposed corpus. Section 4 details the deep learning models applied to our proposed corpus. Section 5 shows the results and their analysis. Finally, Section 6 concludes the research presented in the article.

2. Related Work

Research on emotion detection started in the last few decades (Canales and Martinez 2014). The majority of work for emotion detection concentrated on English and other international languages (Chatterjee, et al. 2019) (Baali and Ghneim 2019) (Syed and others 2015), and a few attempts for Urdu (Bashir, et al. 2022), Arabic (Baali and Ghneim 2019), and Indonesian (Arifin, et al. 2014) languages. This section provides a review of principal literature on emotion detection tasks for different languages, including Urdu.

EmoContext presented in SemEval-19 task 3 is a significant contribution to contextual emotion detection for the English language (Chatterjee, et al. 2019).[1] A public dataset was developed containing 38k labeled dialogues distributed over 115k sentences. All dialogues were classified into four emotion classes, i.e., sad, happy, angry, and others. Although diverse approaches had been designed and experimented on this corpus, Bidirectional LSTM based models showed the highest result. The highest achieved micro-averaged F_1 score was 79.59.

Another prominent contribution to English emotion detection is (Ghosh, et al. 2020). A dataset of 18,921 labeled tweets distributed over six basic emotions, namely, happiness, anger, sadness, disgust, surprise, and fear. In the conducted experiments, the authors split training, validation, and testing sets containing 70%, 20%, and 10% tweets respectively. On this dataset, CNN, Bi-GRU, Bi-LSTM, and HAtED models were applied. By implementing the HAtED model, an accuracy of 81% was achieved.

There has been a proposal for emotion detection in the Chinese language (Lai, et al. 2020). A dataset consisting of 15,664 microblogs was prepared and divided evenly over seven emotion classes i.e., happiness, disgust, like, anger, fear, sadness, and surprise. Moreover, the authors applied a syntax-based graph convolution network (GCN) model on a given dataset. Overall, an 82.32% F_1 score was achieved.

Regarding Arabic emotion detection, an important research contribution is presented in (Baali and Ghneim 2019).[2] A dataset containing 5,600 tweets was split into two, with 5,064 tweets as the training dataset and 561 tweets as the testing dataset. The emotions were distributed over four classes sadness, fear, anger, and joy, where there were 1,400 tweets for each class. SVM, Naïve Bayes, Multi-Layer Perceptron, and CNN-based models were developed, and trained and tested on the dataset. The highest accuracy was of 99.82%, achieved for the CNN-DNN-based model.

[1] https://aka.ms/emocontextdata Last visited: 19-Oct-2022

[2] https://competitions.codalab.org/competitions/17751#learn_the_details-datasets Last visited: 19-Oct-2022

Hindi-language centering emotion detection effort is made in (Vijay, et al. 2018). This work contributed to Hindi-English parallel languages. A corpus containing 2,866 sentences, categorized into fear, disgust, surprise, happiness, sadness, and anger, containing 85, 291, 182, 595, 878, and 667 sentences respectively. In this dataset, 168 sentences have multiple emotion classes. The highest accuracy 58.2 is reported and is achieved using the SVM classifier.

For the Urdu language in literature, we found two significant contributions (1) sentiment-annotated lexicon (Syed, et al. 2010) and (2) RUED corpus (Arshad, et al. 2019). In the former contribution, researchers constructed a corpus of 753 processed reviews, containing 361 and 392 positive and negative reviews respectively, designed the SentiUnits-based method, and reported a maximum of 78% accuracy. While in the latter work, a corpus named "RUED" containing 10,000 Roman-Urdu-based sentences was developed. It was annotated using four classes, Happy, Sad, Angry, and Neutral, however, no experimental results were reported.

Recent advancement in the Urdu Emotion detection task were presented in (Bashir, et al. 2022). The authors created a corpus called the Urdu Nastalique Emotions Dataset (UNED), which contains 52k labeled sentences. Furthermore, six emotion classes were used for tagging: happy, sad, in love, angry, afraid, and neutral. The authors used word embedding, BOW, and TF-IDF approaches for feature extraction. Finally deep learning models were applied to this dataset for classification using LSTM and achieved 85% accuracy.

Overall, as far as we observed, the first limitation of existing research is that there is a lack of contextual text emotion detection corpus and techniques for Urdu. Furthermore, the research community largely focused on designing traditional machine learning approaches (Naïve Bayes, SVM, Multi-layered Perceptron) and there have been few attempts at crafting deep learning approaches for Urdu language processing tasks.

3. Corpus Generation Process

a. Source Corpus

Our proposed CUTEC corpus was created using EmoContext: Contextual Emotion detection corpus. EmoContext was originally constructed by Microsoft for an English emotion detection task and was presented in SemEval-2019 Task 3 (Chatterjee, et al. 2019). Overall, the corpus contained 38,424 dialogues and 115,272 sentences. The corpus was released with single training and two testing sets (Test 1 and Test 2). Test 1 carried 2,755 dialogues, and Test 2 contained 5,509 dialogues. Moreover, in Test 2 all dialogues were labeled, while in Test 1 all dialogues were unlabeled. Human experts assigned a label to the entire training and testing sets. Fleiss' Kappa score of the source corpus was 0.58 on training and 0.59 on testing sets, respectively. Each instance (dialogue) is composed of three sentences (turn-1, 2, and 3) to capture contextual information. Turn-1 included the utterance of a user, turn-2 carried the responsibility of a bot, and turn-3 again included the utterance of the user. The outputs were characterized into four labels, i.e., happy, angry, sad, and others. To ensure the quality of the labels of the corpus, formally, 50 judges were trained. Later, 7 judges assigned the final class. The emotionless sentences in the corpus were tagged as "others". The corpus was cleaned by following some important steps: (1) the personally identifiable Information (PII) such as (name, emails, and phone numbers) was removed from the corpus, (2) non-English conversations were removed, (3) offensive conversations were filtered, such as ethnic-religious content, reference to violence, crime or illegal activities. The corpus is freely available for research purposes under Creative Commons Attribution 4.0.[3]

[3] https://aka.ms/emocontextdata Last visited: 19-Oct-2022

b. Translation via Google Translator

To develop our proposed Contextual Urdu Textual Emotion Corpus (CUTEC), the entire English text (omitting Test 1) of the source corpus was translated into Urdu using a world-widely recognized translator, i.e., Google Translator. Google Translate is a free multilingual neural machine translation service developed by Google, to translate text and websites from one language into another.[4] The reason for the suspension of Test 1 is that the instances were not labeled. For manual translation, each sentence was chosen from the dialogue (each dialogue was composed of three sentences) and placed as input text in Google Translator and the output was noted in a plain text file. As a whole, 106,527 sentences were manually translated into Urdu.

c. Validation of Translation

i. Sentence-wise invalid translation

Various sentences in the EmoContext corpus contained shorthand words, e.g., "plz", "enaf", "M8". In one scenario "Enough" was written as "enaf". A human can easily consider these two words alike, however, Google Translator was not able to recognize and translate them. To address this problem, (1) we manually added the correct translation and replaced shorthand words with their full form. For instance, by replacing "plz" with "Please" Google Translator produced the correct translation. Table 1. contains some examples chosen from the CUTEC corpus regarding sentence-wise invalid translation. This table holds the unique number for each instance in CUTEC, the turn of the sentence, the sentence as input to the translator, invalid output by the translator, and the correct translation of input.

Table 1. Sentence-wise Invalid Translations

Id	Turn	Sentence	Translator's output	Correct translation
111	Turn 1	Plz darling 😭😭😭😭😭	Plz darling😭😭😭😭😭	براہ کرم پیاری 😭😭😭
190	Turn 1	That's enaf for me	ہے enaf یہ میرے لئے	یہ میرے لئے کافی ہے
151	Turn 2	Looks like a ficus.	.کی طرح لگتا ہے ficus ایک	انجیر کے درخت کی طرح لگتا ہے۔
168	Turn 2	I'll just stick with my M8. I see no benefit to upgrading.	کے ساتھ رہوں گا۔ M8 میں صرف اپنے مجھے اپ گریڈ کرنے میں کوئی فائدہ نہیں ہے۔	میں صرف اپنے میٹ کے ساتھ رہوں گا۔ مجھے اپ گریڈ کرنے میں کوئی فائدہ نہیں ہے۔
124	Turn 3	I'm not fine	؛ ٹھیک نہیں ہوں&apos میں	میں ٹھیک نہیں ہوں
174	Turn 3	I wish WE could hangout.	کر سکتے۔ hangout کاش ہم	کاش ہم مل کر گھوم سکتے ہیں۔

[4] https://translate.google.com/ Last visited: 19-Oct-2022

ii. Context-wise invalid translation

Another problem that occurred during the development of the dataset was contextually invalid translation. Google Translator, converted sentences based on words only, rarely focusing on context. For example, the sentence “gives you a patient smile I know, love.” was translated into “مجھے معلوم ہے ، پیار ، آپ کو مریض کی مسکراہٹ دیتا ہے۔” if re-translate the Urdu sentence into English, “I know that, Love, gives you patient (medical patient, medical case) smile.” which are literal meanings but not contextual. Google Translator considered the word “patient” as “medical case (medical patient)” but not as “calm, serene, or tolerant”. In this scenario, the English sentences were correct, but the Urdu translation was contextually wrong. Linguistic experts solved these cases by writing the correct translation, later that translation was added to CUTEC. Table 2 contains further examples of such cases.

Table 2. Context-wise Invalid Translation

Id	Turn	Sentence	Translator’s output	Correct translation
229	Turn 2	gives you a patient smile I know, love.	مجھے معلوم ہے ، پیار ، آپ کو مریض کی مسکراہٹ دیتا ہے۔	میں جانتا ہوں کہ آپ کو ایک خوش کن مسکراہٹ ملتی ہے۔ محبت.
252	Turn 1	Really of course	واقعی	واقعی ، یقینا
1089	Turn 3	Open your arms to cuddle.	اپنے بازوؤں کو کھونے کے ل. کھولیں۔	گلے ملنے کے لئے اپنے بازو کھولیں۔

iii. Emoji Validation

Generally, sentences containing emojis are more emotionally representational and bolder. During the development of the dataset, sentences containing emojis are sometimes translated incorrectly, or emojis are not present in translation. As an example, the translation of “please darling😭😭😭😭😭 ” was given as “براہ کرم پیاری”, neglecting emojis. However, “😭😭😭 براہ کرم پیاری” is the correct translation. This problem is solved using a space as a separator between English sentences and emojis, after space Google Translator identified emojis, and placed them accurately in translation.

iv. Label Validation

Class labeling is a supreme activity in the development of supervised learning methods. We selected EmoContext as an input corpus in which all instances had been labeled by 7 human judges. An interesting point observed in this study is that when we translate a sentence from English to Urdu, the emotions enclosed in the message remain the same. For instance, the English sentence “My friend is dead” is translated into Urdu as “میرا دوست فوت ہوگیا“. However, both messages consolidate the same emotion class, that is, “sad”. Thus, we translated all sentences, one by one, from EmoContext to CUTEC and class labels are the same in both corpora.

EmoContext corpus is translated with context and the sequence of turns of instances, so the validity of labels remains the same. Nevertheless, labels were reconfirmed by three linguistic experts. Both English and Urdu instances were provided to experts, and they identified the context of dialogues identically and applied the same label to Urdu dialogues.

d. Standardization of Corpus

CUTEC has been developed as a standard corpus, the corpus is available in a plain text file.[5] Each instance contains 3 sentences to cover contextual information, following the same standards used in the source corpus. The corpus will be released publicly after the publication of the paper under the license of Creative Commons Attribution 4.0 International license.

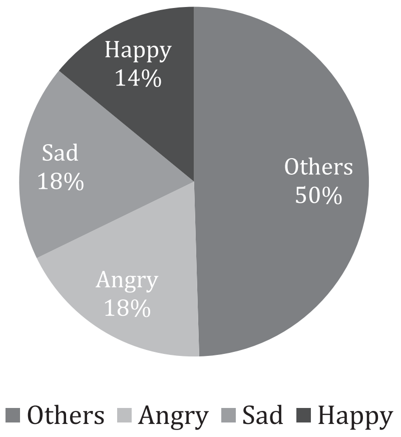


Figure 1. Class-wise Division of Training Dataset in CUTEC

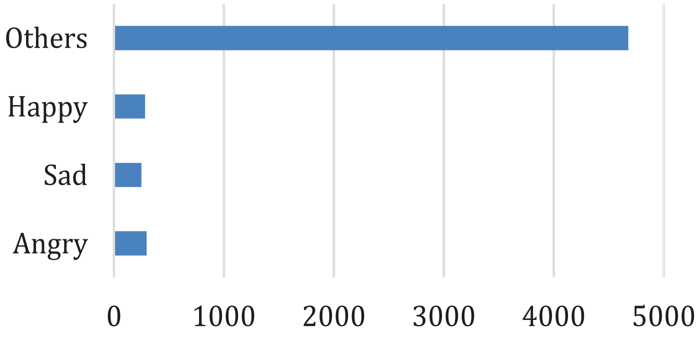


Figure 2. Division of Classes in Testing Dataset

e. Characteristics of Corpus

CUTEC (Contextual Urdu Textual Emotion Corpus) contains 606,548 words in total and 35, 509 labeled dialogues. It is divided into 30,000 instances (90,000 sentences), leading to 530, 381 words as the Training dataset. While the Testing dataset contains 5,509 instances (16,524 sentences), which

[5] https://comsatsnlpgroup.wordpress.com/ Last Visited 11-Nov-2022

amounts to 76,167 words. The credibility of the dataset remains as the source corpus because translation only converts words whereas context remains the same. Figure 1 illustrates the distribution of classes in the training dataset. Overall, 50% of instances belong to the class others, 18% to the class Anger, 18% to Sad, and 14% to Happy. Figure 2 contains the distribution of classes in the test dataset, where "Others" is labeled as 85% of the testing dataset.

4. Experimental Setup

a. Most Frequent Class

Most Frequent Class also referred to as Majority Class classification is a simple approach used by a large number of researchers as a baseline approach to an emotion detection system (Inkpen, Keshtkar and Ghazi 2009) (Krcadinac, et al. 2013). This approach works on a simple phenomenon that assigns the majority class to all instances in the testing set and calculates the percentage of accuracy. Table 5 contains the result of this approach, that is, 84.89%.

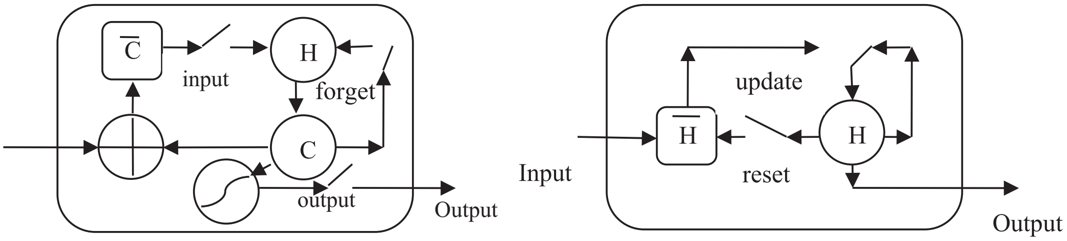
b. Model Description

Figure 4 describes the deep neural network-based architecture used for the Urdu Emotion Detection task. As the figure shows, our model inputs labeled instances from a newly developed corpus (CUTEC). A single instance is composed of three sentences. Further, this model concatenated all three sentences as a single sentence and used it with the class label to train and test deep learning models. Averaged results in terms of Accuracy, Precision, Recall, and F_1 are presented in Table 5.

Furthermore, all concatenated instances (a combination of three sentences) are used as input to the Embedding layer[6]. The purpose of the Embedding layer is to turn the input (words) into low-dimensional vector representations (Fang, et al. 2016) (Hochreiter and Shmidhuber 1997). These Embeddings can conveniently be used to develop embedding-based features. This study used embedding-based features to develop various deep neural network approaches.

Overall, we designed three-layered based models, (1) a single hidden layer, (2) two hidden layers, and (3) three hidden layers. The single hidden layer used either one type of processing unit at one time, i.e., Simple Recurrent Neural Networks (SRNNs) (Bullinaria 2013), Long Short-Term Memory (LSTM) (Hochreiter and Shmidhuber 1997) (Figure 3. (a) Shows an LSTM unit), Bidirectional Long Short-Term Memory (Bi-LSTM) (Zhang, et al. 2015), Gated Recurrent Units (GRUs) (Abdul, Mageed and Ungar 2017) (Figure 3. (b) Shows GRU Unit) and Bidirectional Gated Recurrent Units (Bi-GRUs) (Yu, Zhao and Wang 2019). The reason for using Recurrent Neural Networks based approaches is well suited for sequence learning and text processing tasks (Baali and Ghneim 2019) (Bashir, et al. 2022). These models return the state-of-the-art performance for different NLP tasks (Howard and Ruder 2018).

[6] We used Keras (https://keras.io/) to implement Deep Learning models. According to the specifications of Keras, the embedding layer can only be used as a first layer of the model as we did in this study.



(a) Long-Short-Term Memory Unit

(b) Gated-Recurrent-Unit

Figure 3. The Architecture of the Basic LSTM Unit (obtained from (Hochreiter and Shmidhuber 1997)) and GRU Unit (obtained from (Zhao, et al. 2019))

Figure 3 (a) illustrates that LSTM is composed of three gates, namely, input, output, and forget. Figure 3 (b) GRU is comprised of a comparatively simple architecture with two gates rest and update. C and C illustrate previous and upcoming memory cell contents. Whereas H and H are the candidate activation functions.

After using different types of processing units added on the hidden layer(s), we added a dropout layer to the model. Dropout is a regularization technique used to reduce the overfitting of the model. A dropout value of 0.2 (20%) is widely used to avoid overfitting (Abdullah, Hadzikadicy and Shaikhz 2018) (Krishna and Patil 2020). Lastly, on the output layer, Sigmoid function is used to classify each dialogue concerning its emotion class (Wang and others 2020) (a numeric representation of each class is available in Table 3).

Table 3. Corpus Classes and Respective Numeric Label

Class	**Angry**	**Sad**	**Happy**	**Others**
Numeric Label	0	1	2	3

Table 4 represents the parametric values used by deep neural networks for the textual emotion detection task. Experiments were conducted using three different hidden layers (from 1 to 3), a fixed number of 100 neurons in each hidden layer, Tanh as activation function, applied a constant 0.001 learning rate, batch size of 32, validation split as 20% and persistent dropout value of 0.2 (20%). The output layer used four specified neurons with a Sigmoid activation function. This study used an optimal value of epochs, whereas a large number of epochs can cause an overfitting problem, however, small numbers may lead to an under-fit model (Zhang, Zhang and Jiang 2019). During experiments, observed different values of the number of epochs (ranges from 5 to 40), and found 20 as the most optimal value.

Table 4. Parametric Configurations for All Experiments

Learning Rate	**Default**
Number of Epochs	20
Batch Size	32
Activation Function in Hidden Layers	Tanh
Activation Function in Output Layer	Sigmoid

(continued)

Table 4. Parametric Configurations for All Experiments (continued)

Learning Rate	**Default**
Embedding	1000
Number of Neurons in Hidden Layer	100
Number of Neurons in Output Layer	4
Validation Split	0.2
Dropout	0.2

c. Evaluation Methodology

In this study, the contextual emotion detection task has been studied as multi-class classification. The model has to distinguish four classes, i.e., happy, sad, angry, and others. CUTEC has been used for training and testing deep neural network-based models. In CUTEC, data had already been split into two sets, that is, there were a total of 30,000 training (Figure 1 shows the class-wise detail of training instances) and 5,509 testing instances (Figure 2 shows the class-wise detail of testing instances). This study used the entire data for experimentations and reported results in Table 5. Furthermore, this study used the Keras embedding layer for LSTM, GRU, Bi-LSTM, and Bi-GRU models in the input layer.

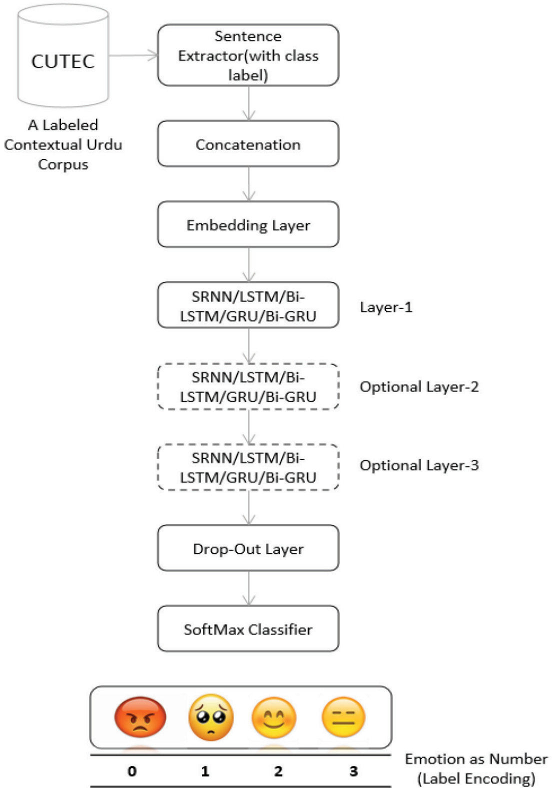

CUTEC
A Labeled Contextual Urdu Corpus
Sentence Extractor(with class label)
Concatenation
Embedding Layer
SRNN/LSTM/Bi-LSTM/GRU/Bi-GRU
Layer-1
SRNN/LSTM/Bi-LSTM/GRU/Bi-GRU
Optional Layer-2
SRNN/LSTM/Bi-LSTM/GRU/Bi-GRU
Optional Layer-3
Drop-Out Layer
SoftMax Classifier
0
1
2
3
Emotion as Number (Label Encoding)

Figure 4. The Architecture of the Deep Learning Model Used for the Urdu Emotion Detection Task.

d. Evaluation Measure

In this study, the evaluation has been carried out using the widely applied evaluation measures i.e. Accuracy, Precision, Recall, and F$_1$ (Bashir, et al. 2022) (Abdullah, Hadzikadicy and Shaikhz 2018) (Acheampong, Wenyu and Nunoo-Mensah 2020). These measures are largely used by the research community to evaluate the performance of emotion detection systems (Acheampong, Wenyu and Nunoo-Mensah 2020). The Following equations have been used to calculate the measures, which are obtained from (Saeed, et al. 2019) (Al-Saqqa, Abdel-Nabi and Awajan 2018).

The Accuracy of a system is defined as the total number of correct predictions.

$$Accuracy = \frac{CorrectCases}{Allcases} \times 100 \tag{1}$$

The precision of a system is the ability of the classifier not to call a negative sample positive.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{2}$$

The Recall of a system is the ability of a classifier to find all the positive samples.

$$Recall = \frac{TruePositive}{TruePpsitive + FalsNegative} \tag{3}$$

F$_1$ measure is a specific relationship (harmonic mean) between precision (Pre) and recall (Rec).

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Accuracy, Precision, Recall, and F$_1$ scores have been computed using all the 5.509k testing dataset and reported in this study.

5. Results and Analysis

Table 5 shows the results in terms of Accuracy, Precision, Recall, and F$_1$ obtained using five deep learning methods, i.e., LSTM, Bi-LSTM, GRU, Bi-GRU, and Simple RNN with three-layer based approaches, i.e., 1, 2 and 3 hidden layers.

Overall, on our proposed corpus CUTEC, the highest results (Accuracy = 87.28% and F$_1$= 0.87) have been achieved using the GRU based model, and the lowest results (Accuracy = 66.62% and F$_1$= 0.72) have been attained using the LSTM-based model. The results highlight the fact that the models that used two hidden layers are the most appropriate for Urdu contextual emotion detection task. The possible reason is that the training data that have been used for experimentation were of a moderate size (approximately 30K instances). To train the neural model with three or more layers, more training data is required which is unfortunately not available for the Urdu language. An important conclusion can be drawn from the results, namely, that a large number of deep learning methods showed results higher than the baseline approach (Accuracy = 84.89%). Further, this study presents a more accurate model (2-layers GRU architecture) in comparison to the previous studies which had carried out the same task (Bashir, et al. 2022) (Rehman and Bajwa 2016).

Table 6 compares the features of our proposed corpus with existing corpora for the textual emotion detection task. As can be noted from the above table our proposed corpus (CUTEC) has more tagged sentences (106, 527 sentences) compare to the existing corpora. Furthermore, we used 4 classes, which are widely used for textual emotion detection task (Arshad, et al. 2019).

Table 5. Results obtained using different deep learning modes for contextual Urdu emotion detection task

No of Layers	Corpus	Model	Accuracy	F_1-Score	Precision	Recall
1	CUTEC	LSTM	66.62%	0.72	0.72	0.67
		Bi-LSTM	82.12%	0.83	0.83	0.82
		GRU	85.55%	0.86	0.86	0.86
		Bi-GRU	86.60%	0.87	0.87	0.87
		Simple RNN	67.51%	0.72	0.72	0.68
2	CUTEC	LSTM	68.22%	0.73	0.79	0.74
		Bi-LSTM	67.53%	0.72	0.80	0.68
		GRU	**87.28%**	**0.87**	**0.88**	**0.87**
		Bi-GRU	86.15%	0.86	0.87	0.86
		Simple RNN	73.95%	0.76	0.80	0.74
3	CUTEC	LSTM	79.31%	0.81	0.85	0.79
		Bi-LSTM	81.21%	0.83	0.85	0.81
		GRU	85.02%	0.86	0.87	0.85
		Bi-GRU	83.92%	0.85	0.87	0.84
		Simple RNN	84.90%	0.78	0.72	0.85
(Bashir, et al. 2022)	UNED	Bi-LSTM	85.30%	0.84	0.87	0.85
(Rehman and Bajwa 2016)	Urdu-Lexicon	Text Similarity	66.00%	0.73	0.69	0.79
MFC (Baseline)		-	84.89%	-	-	-

Table 6. Comparison of CUTEC with Existing Corpora

Corpus	Language	Emotion Classes	Labeled Sentences
UNED (Bashir, et al. 2022)	Urdu	6	52,000 sentences and 2,000 paragraphs
RUED (Arshad, et al. 2019)	Roman Urdu	4	10,000 sentences
Urdu-Lexicon (Rehman and Bajwa 2016)	Urdu	3	-
CUTEC (current study)	**Urdu**	**4**	**35, 509 x 3 = 106, 527 sentences**

Table 7 describes the class-wise detail of F_1 achieved using five deep-learning models. The model that shows the top results, i.e., GRU with two hidden layers is stronger on "Angry and Sad" than "Sad and Happy" in terms of F_1 score. Further, LSTM and Simple are not well suited for Urdu contextual emotion detection task.

Table 7. Class-wise Detail of F_1-score Results

No of Layers	Model	Angry	Sad	Happy	Others
1	LSTM	0.28	0.24	0.23	0.80
	Bi-LSTM	0.54	0.46	0.41	0.90
	GRU	0.58	0.52	0.53	0.92
	Bi-GRU	0.59	0.53	0.53	0.92
	Simple RNN	0.38	0.22	0.16	0.80
2	LSTM	0.27	0.26	0.16	0.81
	Bi-LSTM	0.35	0.20	0.25	0.80
	GRU	**0.61**	**0.54**	**0.56**	**0.93**
	Bi-GRU	0.58	0.54	0.48	0.92
	Simple RNN	0.33	0.26	0.18	0.85
3	LSTM	0.47	0.46	0.42	0.88
	Bi-LSTM	0.54	0.48	0.40	0.89
	GRU	0.59	0.53	0.51	0.91
	Bi-GRU	0.58	0.50	0.51	0.91
	Simple RNN	0.00	0.00	0.00	0.92

Table 8 presents class-wise detail of Precision computed using different deep-learning models. The results show that in terms of Precision, for the class "Others" the most appropriate methods are that usually use three hidden layers.

Table 8. Class-wise Detail of Precision Results

No of Layers	Model	Angry	Sad	Happy	Others
1	LSTM	0.20	0.17	0.19	0.90
	Bi-LSTM	0.45	0.39	0.36	0.93
	GRU	0.51	0.45	0.55	**0.94**
	Bi-GRU	0.56	**0.51**	0.52	0.93
	Simple RNN	0.29	0.15	0.13	0.90
2	LSTM	0.28	0.14	0.19	0.90
	Bi-LSTM	0.35	0.20	0.25	0.80
	GRU	**0.58**	0.48	**0.63**	0.93
	Bi-GRU	0.52	0.48	0.58	0.93
	Simple RNN	0.23	0.20	0.30	0.90
3	LSTM	0.35	0.38	0.35	**0.94**
	Bi-LSTM	0.44	0.39	0.34	0.93
	GRU	0.52	0.47	0.45	**0.94**
	Bi-GRU	0.51	0.40	0.47	**0.94**
	Simple RNN	0.00	0.00	0.00	0.85

For classes "Happy" and "Angry" the most suitable method is GRU with two hidden layers in terms of Precision. For class "Sad" Bi-GRU shows the highest results in term of Precision.

Table 9. illustrates the Recall of models. Where the highest recall over the "Angry" class is achieved by LSTM-3. Bi-GRU-3 performed highest over the "Sad" Class. The "Happy" class's maximum recall of 0.59 was achieved by GRU-3, and the highest recall over "Others" was achieved by SimpleRNN-3 which was 1.0. if neglect SimpleRNN-3 then by Bi-GRU-1, Bi-GRU-2, GRU-2 is 0.92.

Table 9. Class-wise Detail of Recall Results

No of Layers	Model	Angry	Sad	Happy	Others
1	LSTM	0.48	0.42	0.29	0.72
	Bi-LSTM	0.67	0.56	0.47	0.87
	GRU	0.69	0.61	0.50	0.90
	Bi-GRU	0.63	0.56	0.54	0.92
	Simple RNN	0.56	0.42	0.22	0.72
2	LSTM	0.43	0.42	0.21	0.27
	Bi-LSTM	0.48	0.40	0.35	0.72
	GRU	0.63	0.63	0.51	0.92
	Bi-GRU	0.64	0.63	0.40	0.92
	Simple RNN	0.58	0.36	0.13	0.81
3	LSTM	**0.71**	0.58	0.51	0.83
	Bi-LSTM	0.69	0.62	0.47	0.84
	GRU	0.70	0.61	**0.59**	0.89
	Bi-GRU	0.67	**0.69**	0.56	0.87
	Simple RNN	0.00	0.00	0.00	**1.00**

6. Conclusion

Urdu is an under-resourced language with a significant number of speakers. Urdu requires labeled corpora to enhance different NLP tasks. As presented in the paper, the contribution of this study is a newly constructed and publicly available benchmark corpus called CUTEC. The corpus contains 35,509 labeled dialogues, where each dialogue is composed of three Urdu sentences. In addition to the construction of the corpus, this study experimented with five widely used deep learning models to check their stability. The results of our contextual emotion detection task show that GRU achieves the best performance. In the future, we are interested in studying transfer learning, ensemble learning, and generative adversarial networks for the task of emotion detection in Urdu.

References

Abdul, M., M., and Lyle, U., 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. Proceedings of the 55th annual meeting of the association for computational linguistics, Vancouvre, Canada, ACL, 1, 718-728.

Abdullah, M., Mirsad, H., and Samira, S., 2018. SEDAT: sentiment and emotion detection in Arabic text using CNN-LSTM deep learning. 17th IEEE international conference on machine learning and applications (ICMLA), Florida, USA, IEEE, 5-840.

Acheampong, F. A., Chen, W., and Henry N. M., 2020. Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2(7), e12189.

Al-Saqqa, S., Heba, A., N., and Arafat, A., 2018. A survey of textual emotion detection. 8th International Conference on Computer Science and Information Technology (CSIT). Amman, Jordan, 136-142.

Arifin, A., Z., Yuita A., S., Evy K., R., and Siti M., 2014. Emotion Detecion of Tweets in Indonesian Language using Non-Negative Matrix Factorization. International Journal of Intelligent Systems and Applications 6(9), 54.

Arshad, M., U., Muhammad, F., B., Adil, M., Waseem, S., and Mirza, O., Beg., 2019. Corpus for emotion detection on roman urdu. 22nd International Multitopic Conference (INMIC). Islamabad, Pakistan, 1-6.

Ayir, A., Iil, Y., and Hasan, D., 2018. Feature extraction based on deep learning for some traditional machine learning methods. 3rd International Conference on Computer Science and Engineering (UBMK), USA, 494-497.

Baali, M., and Nada, G., 2019. Emotion analysis of Arabic tweets using deep learning approach. Journal of Big Data. 6(1), 1-12.

Bashir, M., F., Abdul R., J., Muhammad U., A., Thippa R., G., Waseem S., and Mirza O., B., 2022. Context aware emotion detection from low resource urdu language using deep neural network. Transactions on Asian and Low-Resource Language Information Processing, 2022.

Bullinaria, J., A., 2013. Recurrent neural networks. Neural Computation: Lecture 12.

Canales, L., and Barco, P., M., 2014. Emotion detection from text: A survey. Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC), Quito, Ecuador, 37-43.

Chang, C., and Michael, M. 2020. Using word order in political text classification with long short-term memory models. Political Analysis 28(3), 395--411.

Chang, V., 2016. Review and discussion: E-learning for academia and industry. International Journal of Information Management 36(3), 476--485.

Chatterjee, A., Kedhar, N., N., Meghana, J., and Puneet, A., 2019. SemEval-2019 Task 3: EmoContextContextual Emotion Detection in Text. International Workshop on Semantic Evaluation. Minneapolis: MIT press, 39-48.

Druck, G., and Bo, P., 2012. Spice it up? Mining refinements to online instructions from user generated content. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jejo Island, Korea, 545-535.

Fang, W., Jianwen, Z., Dilin, W., Zheng, C., and Ming, L., 2016. Entity disambiguation by knowledge and text jointly embedding. Proceedings of the 20th SIGNLL conference on computational natural language learning. Berlin, Germany, 260-269.

Ghosh, Soumitra, et al. 2020. Annotated Corpus of Tweets in English from Various Domains for Emotion Detection. Proceedings of the 17th International Conference on Natural Language Processing (ICON). Patna, India, 460-469.

Hochreiter, S., and Jurgen, S. 1997. Long Short-Term Memory. Neural Computation 9(8), 1735-1780.

Howard, J., and Sebastian, R., 2018. Universal Language Model Fine-tuning for Text Classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 328-339.

Hussain, S., 2008. Resources for Urdu language processing. Proceedings of the 6th workshop on Asian Language Resources.

Inkpen, D., Fazel, K., and Diman, G., 2009. Analysis and generation of emotion in texts. KEPT, 3--13.

Kao, E., C., C., Chun-Chieh, L., Ting-Hao, Y., Chang-Tai, H., and Von-Wun, S., 2009. Towards text-based emotion detection a survey and possible improvements. International Conference on Information Management and Engineering, New Zeeland, IEEE, 70-74.

Khan, W., Ali D., Jamal, A N., and Tehmina, A., 2016. A survey on the state-of-the-art machine learning models in the context of NLP. Kuwait journal of Science. 43(4).

Krcadinac, U., Philippe, P., Jelena, J., and Vladan, D., 2013. Synesketch: An open source library for sentence-based emotion recognition. IEEE Transactions on Affective Computing. 4(13). 312-325.

Krishna, DN., and Ankita, P., 2020. Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. Interspeech. 4243-4247.

Lai, Y., Linfeng, Z., Donghong, H., Rui, Z., and Guoren, W., 2020. Fine-grained emotion classification of Chinese microblogs based on graph convolution networks. World Wide Web, 23(5), 2771--2787.

Nagwani, N., K., 2015. A comment on “a similarity measure for text classification and clustering”. IEEE Transactions on Knowledge and Data Engineering, 27(9), 2589--2590.

Naseer, A., and Sarmad, H., 2009. Supervised word sense disambiguation for Urdu using Bayesian classification. Center for Research in Urdu Language Processing, Lahore, Pakistan, 2009.

Panko, R., R, and Hazel, G. B., 2002. Monitoring for pornography and sexual harassment. Communications of the ACM, 45(1), 84--87.

Rahman, T., 2004. Language policy and localization in Pakistan: Proposal for a paradigmatic shift. In SCALLA Conference on computational linguistics, Pakistan, 1-19.

Rani, J., and Kanwal G., 2014. Emotion detection using facial expressions-A review. International Journal of Advanced Research in Computer Science and Software Engineering, 4(4).

Rehman, Z., U., and Imran, S., B., 2016. Lexicon-based Sentiment Analysis for Urdu. Sixth international conference on innovative computing technology (INTECH). Dublin, IEEE,497-501.

Riaz, K., 2010. Rule-based named entity recognition in Urdu. Proceedings of the 2010 named entities workshop. Uppsala, Sweden, ACL, 126-135.

Rincon, J., Jose, L., P., Juan, L., P., Vicente, J., and Carlos, C., 2016. Adding real data to detect emotions by means of smart resource artifacts in MAS. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 5(4), 85.

Saeed, A., Rao, M., A., N., Mark, S., and Paul, R., 2019. A word sense disambiguation corpus for Urdu. Language Resources and Evaluation, 53(3), 397--418.

Syed, A. Z, and others. 2015. Applying sentiment and emotion analysis on brand tweets for digital marketing. IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Jordan, IEEE, 1-6.

Syed, A., Z, Muhammad, A., Enriquez, M., and Maria, A., 2010. Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits." Mexican international conference on artificial intelligence. Maxico, IEEE, 32-43

Vijay, D., Aditya, B., Vinay, S., Syed, S., A., and Manish, S., 2018. Corpus creation and emotion prediction for Hindi-English code-mixed social media text. Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: student research workshop, New Orleans, Lousiana, USA, ACL, 128-135.

Wang, Z., and others. 2020, Text emotion detection based on Bi-LSTM network. Academic Journal of Computing \& Information Science, 3(3).

Yu, F., Eric, C., Ying-Qing, X., and Heung-Yeung, S., 2001. Emotion detection from speech to enrich multimedia content. Pacific-Rim Conference on Multimedia, Bejing, China, 550--557.

Yu, Q., Hui, Z., and Zuohua, W., 2019. Attention-based bidirectional gated recurrent unit neural networks for sentiment analysis. Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition, Bejing, China, 116-119.

Zhang, H., Lin, Z., and Yuan, J., 2019. Overfitting and underfitting analysis for deep learning based end-to-end communication systems. 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xian, China, 1-6

Zhang, S., Dequan, Z., Xinchen, H., and Ming, Y., 2015. Bidirectional long short-term memory networks for relation classification. Proceedings of the 29th Pacific Asia conference on language, information and computation, Shanghai, China, 73-78.

Zhao, H., Zhongxin, C., Hao, J., Wenlong, J., Liang, S., and Min, F., 2019. Evaluation of three deep learning models for early crop classification using sentinel-1A imagery time series—A case study in Zhanjiang, China. Remote Sensing, 11(22), 2673.