2024, 21(2 Special Issue): 0725-0739
https://doi.org/10.21123/bsj.2024.9822
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Baghdad Science Journal

Sentiment Analysis on Roman Urdu Students' Feedback Using Enhanced Word Embedding Technique

***Noureen** *[1], **Sharin Hazlin Huspi**[1], **Zafar Ali**[2,3]*

[1]Department of Applied Computing and Artificial Intelligence, Universiti Teknologi Malaysia, Johor, Malaysia.
[2]Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia.
[3]Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan.
*Corresponding Author.

PARS2023: Postgraduate Annual Research Seminars 2023.

Received 02/10/2023, Revised 10/02/2024, Accepted 12/02/2024, Published 25/02/2024

Abstract

Students' feedback is crucial for educational institutions to assess the performance of their teachers, most opinions are expressed in their native language, especially for people in south Asian regions. In Pakistan, people use Roman Urdu to express their reviews, and this applied in the education domain where students used Roman Urdu to express their feedback. It is very time-consuming and labor-intensive process to handle qualitative opinions manually. Additionally, it can be difficult to determine sentence semantics in a text that is written in a colloquial style like Roman Urdu. This study proposes an enhanced word embedding technique and investigates the neural word Embedding (Word2Vec and Glove) to determine which performs better for Roman Urdu Sentiment analysis. Our suggested model employs the BiLSTM network to maintain the context in both directions and eventually, results for ternary classification are obtained by using the final softmax output layer. A manually labeled data set was used to evaluate the model, data is collected from the HEIs of Pakistan. Model was empirically evaluated on two datasets of Roman Urdu, the newly developed student's feedback dataset and RUSA-19 publically available data set of Roman Urdu. Our model performs effectively using the word embedding and BiLSTM layer. The proposed model is compared with the baseline models of CNN, RNN, GRU and classic LSTM. The experimental findings demonstrate the proposed model's efficacy with an F1score of 90%.

Keywords: Long Short-term memory network, Roman Urdu, Sentiment Analysis, Student feedback, Word Embedding.

Introduction

In Pakistan and its subcontinent Roman Urdu is a highly popular language on social media. Roman Urdu is primarily used by peoples to post reviews and other feedback on social media and other venues. One of the dialects of the third-largest Urdu language in the world is Roman Urdu, however further study is required[1]. On social media sites like Facebook, Twitter, Instagram, and Snapchat, millions of Roman Urdu words are uploaded every day. This enormous volume of information has one major flaw: it ignores viewpoints stated in Roman Urdu and only considers the English (resource-rich) language into account. Numerous languages, most notably Hindi and Roman Urdu, are perceived as resource-poor in comparison to English because there aren't as many resources available to conduct sentiment analysis on them[2]. The term "Roman Urdu" refers to the English alphabet used to write the Urdu language in Roman script. Roman Urdu is seen to be an example of a morphologically difficult but academically rich language[3]. Roman Urdu faces multiple challenges because the English and Latin alphabets offer less morphological support. The informal style of Roman Urdu writing is the main obstacle to interpreting the language. To address this issue, researchers have faced a significant barrier due to the uneven representation of text. In other words, there are no norms for Roman Urdu due to spelling irregularities and variations[4]. As the official language of Pakistan, Urdu is used by over 200 million people daily, while another 100 million people worldwide use it to communicate in writing[5]. Less linguistic techniques, including as stop word lists, lemmatization and stemming tools, are available for use in processing language in Urdu. The English alphabet can be used to translate Urdu into Roman script. As an example, the term "Acha" in Roman Urdu means "Good” As a result, the term Acha has many different spellings, like Asha, Achha, Achaa, Ascha etc.

Due to the informal nature of Roman Urdu text, the task of sentiment analysis in natural language processing has always been difficult. Most studies have focused on customer reviews for products, services[5] music, mobile devices, accommodations, etc. and limited work investigated on Education domain especially in the Roman Urdu Language[6]. In existing studies, machine and deep learning techniques have been primarily used to conduct the sentiment analysis task with combination of word embedding. In context of Roman Urdu, standard word embedding is not available because of the colloquial nature of text and the standard feature extraction models are insufficient and must be revised for handling such colloquial text. So, to overcome the limitation of existing studies, this paper represents a first attempt to do sentiment analysis on a corpus of Roman Urdu for faculty teaching evaluations using enhanced word embedding and the BiLSTM model.

The primary contributions this paper makes are as follows:

1. A new student’s feedback dataset (Roman Urdu) has been developed & annotated into three groups: positive, negative, and neutral.
2. Roman Urdu sentiment analysis model has been developed by using BiLSTM and enhanced word embedding technique.
3. The suggested approach is evaluated using various deep neural network configurations (CNN, RNN, GRU, and classic LSTM) to determine which one produces the most accurate enhanced word embedding for Roman Urdu data. The best method for resolving the Sentiment analysis issue in Roman Urdu has been identified by comparing the effectiveness of several deep network configurations.

The rest of the paper is structured as follows: Deep learning and word embedding algorithms for Roman Urdu sentiment analysis are discussed in Section II. The proposed approach is the subject of Section III. Results and conclusions from the experiments are presented in Sections IV and V.

Related Work

Sentiment analysis is a branch of "text classification” and the most important research field in the area of (NLP) natural language processing[7]. The term "sentiment analysis" refers to collecting and analyzing people's opinions, feelings, and perspectives on various issues, like services and products[8]. Sentiment Analysis has achieved greater success in the domain of Products, Restaurants, Services, Movies, and Hotel reviews datasets[5, 9]. Recently, Researchers in the field of education have also given sentiment analysis a lot of attention[10]. Since the COVID-19 epidemic, when the majority of educational institutions switched from traditional face-to-face instruction to an online form, student feedback has grown in popularity and importance. Countries are trying to improve their educational institutions to grow further in society[10]. Higher education institutions are also trying to enhance the quality of education by evaluating instructors' teaching, student progress, and course analysis using feedback[11]. Deep learning (DL) models with the combination of word embedding techniques have significantly enhanced the performance of NLP applications including sentiment analysis and question answering[12-15]. Word embedding is used by Deep Learning models in NLP tasks to automatically extract and recognize high level features by using textual input[16].

In the light of limitations of traditional machine learning approaches, various DNNs have been used for sentiment classification[17]; these include CNNs, RNNs, RAEs, and DBNs. The RNN is the preferred option for sequential modelling tasks, Due to its ability to handle sequences of any length[18]. Similar to the LSTM, the RNN has shown remarkable performance while keeping contextual data to classify texts. After training with enough data and computational power, RNN variants, including LSTM, GRU[19], and Bi-LSTM[20] were utilized for sentiment analysis, with promising results[21, 22]. The author addressed the sentiment analysis problem by comparing Roman Urdu data using deep learning-based approaches. In order to determine which machine learning classifier performed most effectively, the suggested deep learning model, LSTM, was put through its trials. In comparison with machine learning, the deep learning model LSTM was found to be more effective[23]. TF-IDF and GloVe embedding were used with BiLSTM classifier and achieved better results[24].

Table 1. Existing work on Roman Urdu Sentiment Analysis

S.no	Title	Year	Dataset	Algorithm	Feature Extraction	Accuracy
01	"Sentiment Analysis in E-commerce Using SVM on Roman Urdu Text"	2019	E-Commerce Website (Daraz.pk)	SVM	TF-IDF	60.03%
02	"Deep sentiments in Roman Urdu text using Recurrent Convolutional Neural Network mode"	2020	RUSA-19 Corpus & Roman Urdu UCL	RCNN	None	71.3% 69.3%
03	"A Precisely Xtreme-Multi Channel Hybrid Approach for Roman Urdu Sentiment Analysis"	2020	3241 mobile-related Roman Urdu dataset	(RNN, CNN and LSTM)	Word2vec, Glove and FastText	75.75% 75.75% 72.87%
04	"Sentiment Analysis of Roman Urdu on E-Commerce Reviews Using Machine Learning"	2022	Roman Urdu e-commerce dataset (RUECD)	SVM	None	68%
05	"Deep Sentiment Analysis Using CNN-LSTM Architecture of Roman Urdu Text Shared in Social Media"	2022	UCL, RUSA-19	Two layer LSTM	TF-IDF	71.4% 72.6%

Table 1 shows the existing work on Roman Urdu Sentiment Analysis which is briefly summarized here:

In one of the existing study the author proposed the traditional TF-IDF algorithm and BiLSTM is employed to capture context data RNN, CNN, LSTM and NB sentiment analysis techniques were tested. The proposed model BiLSTM has proven to be highly accurate and effective[25]. The Proposed BiLSTM model combined with pre-trained Word2Vec, GloVe and FastText word embedding, for Roman Urdu E-Commerce data, results showed that Word2Vec performed with a reasonable accuracy of 67%[26]. In this study the author evaluates the performance of various word embedding's using conventional machine learning classifiers in conjunction with the CNN-LSTM architecture. The author proposed a deep leaning model for Roman Urdu corpus and presented two-layered sentiment analysis model: LSTM for maintaining contextual information and CNN model layer for extracting features. The author used the datasets of UCL and RUSA-19; experimental results showed that two-layer LSTM classifiers with TF-IDF word embedding performed better with 71.4% and 72.6% accuracy, respectively[27]. The author recently experimented with various ML techniques, and proposed SVM and used the RU e-commerce dataset (RUECD) to evaluate the model's effectiveness. The suggested model used an SVM classifier to achieve 68% accuracy. According to literature review, the SVM performed better than the other machine classifier[28]. RUSA-19 Roman Urdu Corpus was created, which includes 10,012 comments. The reviews were accumulated from numerous social media networks and address a wide range of subjects, consisting of (i) talk shows and drama, movies, (ii) politics, (iii) food recipes, (iv) sports, (v) apps, blogs, and forums; collected from many social media networks. They proposed deep network RCNN. They concluded that the Rule-based model achieved 54% accuracy, and the RCNN model’s performance is better with 71 % accuracy. The recommended deep model used components like recurrent and CNN layers. Many researchers have found that CNN works better with non-sequential data but has not yet shown adequate results when dealing with sequential data[29].

In one of existing research, the author suggested a particularly multi-channel hybrid technique; he concluded that most machine learning techniques for sentiment analysis work better with TF-IDF. However, customized deep-learning techniques work better with word2vec. Experiments showed that the proposed model LSTM with Word2Vec performs better, with an accuracy of 75.5%. It is observed from the existing literature that LSTM and BiLSTM performs better compared to other ML and Deep learning classifiers for the classification of sentiment analysis of RU. It is observed from the existing literature that LSTM and BiLSTM performed better when combined with word embedding technique.

In several NLP problems, although the LSTM has shown to be highly successful, it still has room for improvement. According to the literature, BiLSTM performs better mostly for resource-rich language (English). It is observed from the literature review, limited study has been conducted on (Roman Urdu). So, there is a need to design a sentiment analysis model based on enhanced embedding and BiLSTM for Roman Urdu corpus.

Materials and Methods

Methodology

In this study, an enhanced word embedding method and BiLSTM were used to create a deep learning model for Roman Urdu sentiment analysis. Fig. 1 shows the operational research design for proposed model.

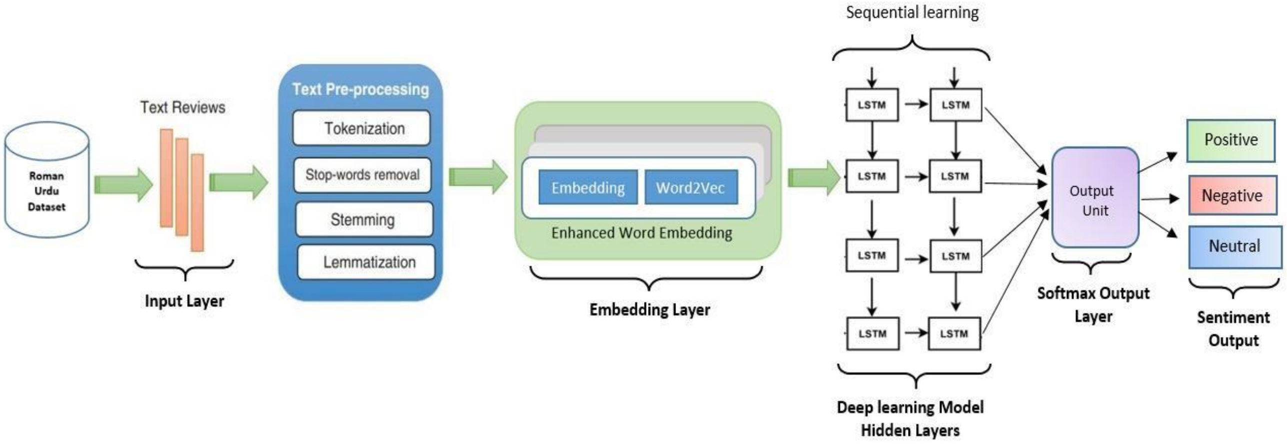

Figure 1. Operational Research Design

The details of the framework, a deep learning model, and the findings will be covered in the next section.

Dataset:

Dataset of student's feedback reviews has been collected for faculty performance evaluation in order to create an academic domain dataset. Data have been collected from Higher education institute (HEIs) of Pakistan. A sample of comments from Roman Urdu dataset with English translation are shown in Table 2.

Table 2. Sample of comments in Roman Urdu with English translation

Roman Urdu Comment	English Translation	Polarity
Hamesha pareshan kerty hen	Always degrades the students	negative
Teacher ko kafee ilm hey	The teacher has vast knowledge	Positive
Class men rawaya bht sakht hey	Rude behavior in the class	negative
Shaandaar teacher	Great Teacher	Positive
Who thoray taiz they	He was a bit speedy	negative
Hamesha students ka khayal kerty hen	Always take care of students	Positive
Parhaany ka tareeqa acha hy	He has good teaching way	Positive
Bas sahi hen	Teacher is just ok	neutral
Unhen parhany ka tareeqa improve kerna chahiye	He should improve his teaching way	neutral
Tadrees ka acha tareeqa	Nice approach of teaching	Positive

Three specialists with a strong grasp of Urdu script manually annotate reviews. Reviews were annotated into three groups (positive, negative, and neutral).

Data Pre-processing:

Basic pre-processing steps like stop-word removal, normalization, and stemming[30] are required before applying the classification to the data. A rule-based lexical normalization approach was used, to normalize the text. Creation of lexical normalization standards has been extremely difficult due to the inconsistent and uneven style of Roman Urdu text. To address this issue, few rules were made, the main goal of the rules is to eliminate known suffixes and infixes from Roman Urdu words. The words such as "buraian" ("badness"), "khamian" ("mistakes"), and "Achaeian" ("goodness"), has been normalized to "khami", "burai" and "achaei" respectively.

Lexical Normalization & Stemming:

Stemming primarily aims to decrease the variety of expressions by combining related words into a single stem word. Before the data is fed in, a stemming procedure is performed using a lexical dictionary annotated by humans. The lexicon based dictionary is developed to standardize textual data with an extensive variety of spellings. In order to group related words with similar meanings, stemming technique was implemented based on dictionaries. Stemming is applied on data after applying the rule based lexical normalization approach. Multiple stemmers are available for English language, including porter-stemmer and snowball stemmer. There isn't a single stemmer available for Roman Urdu, though, it is important that stemming for Roman Urdu is significantly difficult and challenging than the stemming in other languages. A stemming function based on dictionary has been proposed. Table 3 contains an example of lexically normalized terms. It can be seen that words can sound the same phonetically but have various spellings.

Table 3. Table for lexical normalization of Roman Urdu data

Lexicons	Stem Word	English
Achha, Achaa, Asha	Acha	Good
Bohatt, buhatt, boohat	bohat	Very
Behtarr, behtareen,behtarin	behtar	Better
Nahin, nahi, nahe, nhe	nahi	No
Sahee, sahi, sahe, sahii	sahi	Ok

Our stemming method is built on the mapping function *f*: $W \rightarrow S$, which connects the word W to the stem word S. In this case, W is a set of words against which a possible stem word from S is mapped. If the method can't match the particular lexicon to its stem word, it returns the actual word. Further, as shown in Fig. 2, each word's index is separately handled using the hashing technique. Such that the stem word can be immediately retrieved, the abnormalities of data has been reduced by stemming the entire corpus using this mapping strategy.

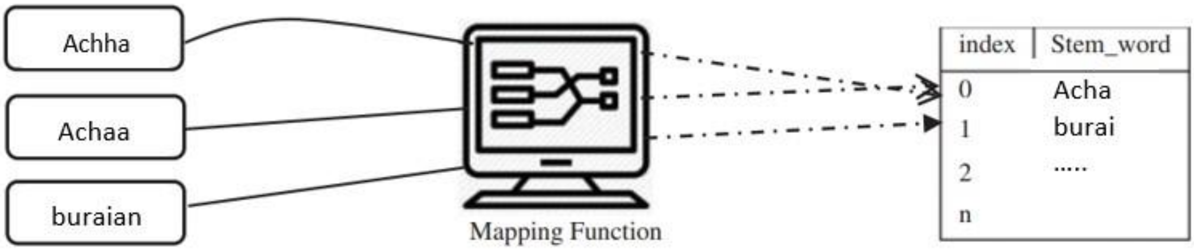

Figure 2. Stemmer mapping function for Roman Urdu

Proposed Model:

The proposed model based on two layer architecture. Layer one represents Enhanced word embedding 2nd layer illustrates (BiLSTM) layer. Fig. 3 describes proposed Roman Urdu Sentiment Analysis model.

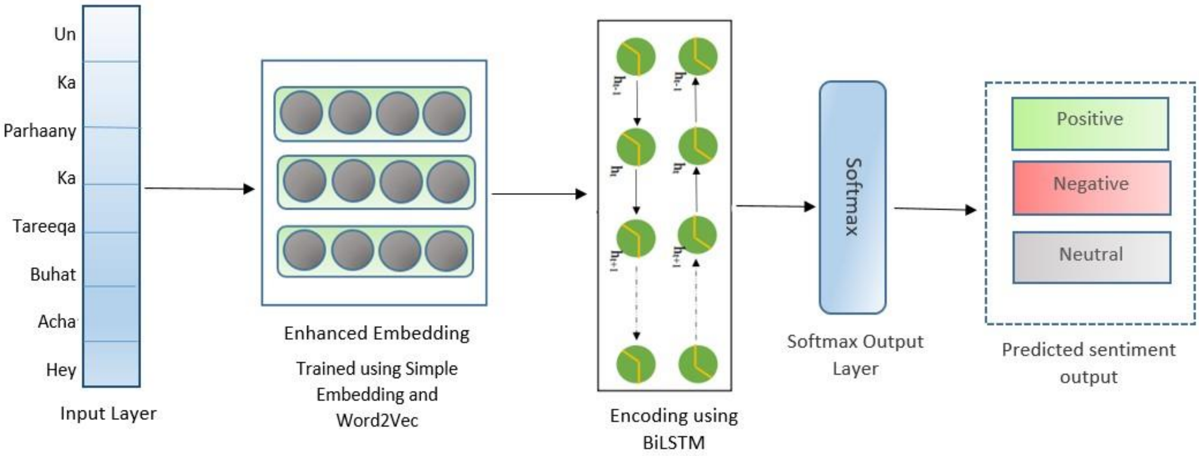

Figure 3. Proposed Roman Urdu Sentiment Analysis Model

The 1st layer represents the word embedding process; In light of the literature review, it is noticed that neural word embedding performs better in the text classification problem; Word embedding is primarily used to generate dense vectors of reduced dimensions for the subsequent layers of a neural network. The colloquial nature of Roman Urdu text is the primary obstacle in processing Roman Urdu. This informal text style has presented researchers with a formidable challenge in addressing this issue. Word embedding's have achieved significant success in performing sentiment analysis tasks, especially in resource-rich languages (English). However, limited studies investigated word embedding for Roman Urdu text because of the informal style of the text. So, our study proposes an enhanced word embedding technique that learned/trained on an annotated data by giving the extra domain-specific knowledge to enhance the quality of word vector. Pre-trained neural embedding Word2Vec and GloVe were employed to investigate which performs better for Roman Urdu Sentiment analysis. At the end, output of enhanced word embedding in the form vectors will be transferred to the neural network layer, the BiLSTM network layer.

In the 2nd layer the proposed BiLSTM model uses bidirectional-LSTM; since it operates with two hidden layers, it processes data in both directions and can intelligently grasp immediate and subsequent contexts. Our data set includes short and long reviews, so our proposed model, BiLSTM can further deal with long-term dependency problem; details are mentioned in next section (Recurrent Neural Network).

Further, the proposed model consists of text sequence input layers: an embedding layer, Bidirectional LSTM layers with specific regularizations; a regularisation dropout layer; RMSprop optimizer, softmax activation function, followed by a dense softmax output layer, which is used to obtain the results of ternary classification.

Recurrent Neural Network:

RNN is capable of collecting previous information. In a nutshell, Simple RNNs are incapable of remembering long time stamps and can only recall short time stamps. This issue is known as long-term dependency, and it can be addressed with LSTM. The LSTM's recurrent architecture can solve the gradient descent's exploding and vanishing problem in addition to maintaining contextual information. LSTM's gating units allow it to regulate the flow and determine what to ignore and update. If there is an input neuron x^t, a hidden output state h^t, and a previous hidden output state $h^{(t-1)}$, simple RNN can be expressed as follows:

$$h^t = \mathrm{g}h(W_i x^t + W_R\, h^{(t-1)} + \ b_h) \qquad 1$$

$$y^t = \mathrm{g}y(W_y\ h^t\ + b_y) \qquad 2$$

Here, h^t represents the hidden output, gh is a squashing function, W is the weight-matrix, b represents a bias, and y^t indicates the final output.

The core idea behind LSTM is that cell states serve as conveyor belts. With little linear interaction, this conveyor belt transfers data in sequences and moves through multiple cell states. During this operation, information is added to or withdrawn from the cell states utilizing gates. The traditional LSTM has four memory blocks, three multiplicative components known as gates, and one cell state c. Fig. 4 depicts the gates as input gate i_t, forget gate f_t, and output gate o_t. A single LSTM memory cell can be demonstrated mathematically as follows:

$$f^t\ = \sigma(W_{xf} x^t + W_{hf}\, h^{t-1} + b_f) \qquad 3$$

$$i^t =\ \sigma(W_{xi} xt + W_{hi} h^t + bi) \qquad 4$$

$$c^{\sim} = \tanh(W_{xc}\, x_t + W_{hc} h^{t-1} + b_c) \qquad 5$$

Update state

$$c_t = f_t \otimes c_t - 1 + i_t \otimes c^{\sim}\, t \qquad 6$$

$$o_t = \sigma(W_{xo} X_T + W_{hc} h^{t-1} + b_o) \qquad 7$$

$$h_t = o_t \otimes \tanh(c_{t)} \qquad 8$$

Here c_t, and h_t represent the cell state, and concealed state, respectively. W is the weight matrix, and b indicates the bias, for each individual layer. First step in the LSTM process is to finding the data that can be eliminated from the cell state as unnecessary. This decision is taken by the forget layer, also known as the sigmoid layer. Second, two-pass layers are used to decide what new information will be added.

Eq.6 states that the previous cell state c_{t-1}, is subsequently changed to the new cell state c_t. In order to identify which part of the cell state could be added as output o_t, a sigmoid layer [0, 1] is eventually executed, it is then inputted with the cell state and element-wise multiplied (abbreviated with) a tanh layer. Values between "1 and 1" are the only ones that the tanh function will accept. This will lead to the production of presented output ht.

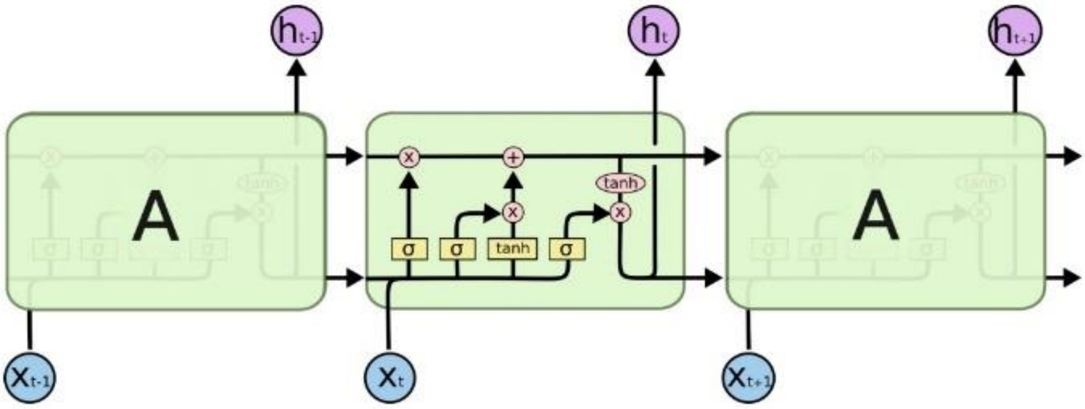

Figure 4. LSTM Main architecture

Learning the Model:

To find the best weights for the suggested model, model learning optimizers were empirically evaluated across the appropriate amount of epochs. For learning the model 27 epochs were used.

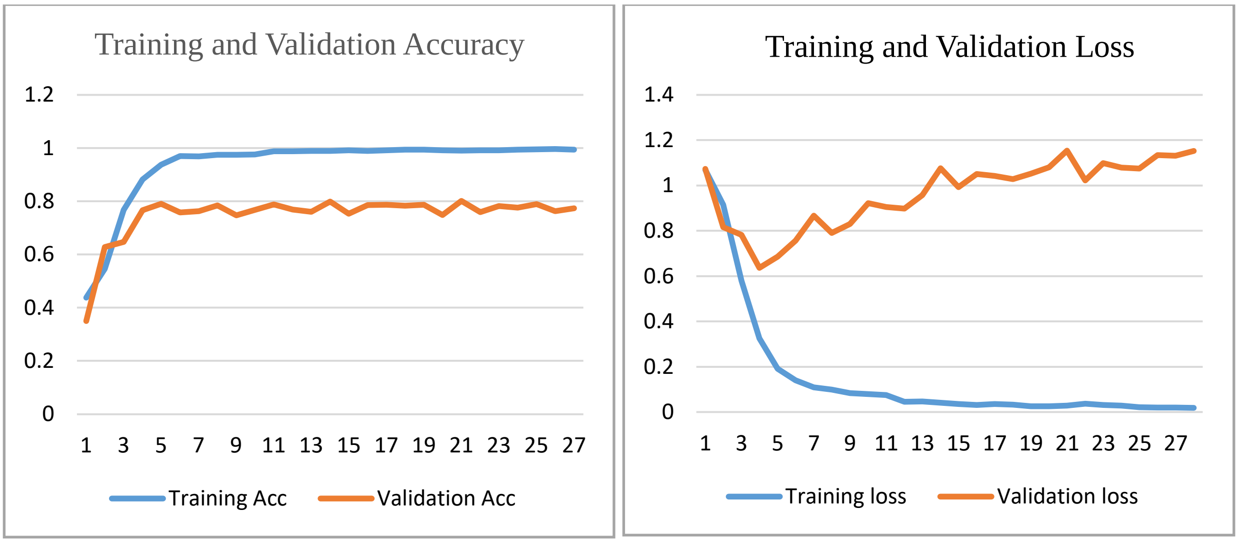

Figure 5. Training and Validation accuracy for ternary classification on the Roman Urdu Student's feedback Dataset

Most deep learning models suffers from the over-fitting issue on the training data, in order to prevent overfitting; drop out layer and batch normalization layer are added. Initially the dropout ratio is set as 0.1 with return sequence set as true, the other drop layer turned to 0.2. When dropout rate was applied more than 0.2 it was observed that model loss the important features and it gets negative impact on the performance. It is clearly shown from Fig. 5 that the accuracy is getting better up to the 27th epochs, and thereafter it generally declines. The experimental evaluation shows that there would be more losses when continue training the model, so final weights that the model learned was after 27 epochs.

Experiments:

Model performance was evaluated on our own created academic domain student feedback Roman Urdu dataset includes (2000 reviews). Model was tested on the well-known RUSA-19[31] (10,021) Customer Reviews publically available dataset in order to see if the proposed model is generalizable. This section describes the experimental design and specific baseline techniques with discussion.

Experimental Setup:

Tensor flow Keras deep learning library has been employed in our test environment. Additionally, the data used for training and testing were separated using sckit-learn. In our empirical configuration the integrity of the neural word embedding's Word2Vec and GloVe will be assessed due to their promising performance in several NLP tasks. Keras Preprocessing deep-learning library is used to generate tokens and apply padding and sequences to balance the size of the tokens.

Table 4. Parameters list for BiLSTM with Word Embedding

Layer Type	Output Shape	Param #
Input_6 (Input Layer)	(None, 877)	0
embedding (Embedding)	(None, 24, 300)	157600
(Bidirectional (bidirectional l)	(None, 24, 48)	62400
Global_max_pooling1d (Globalmaxpooling1D)	(None, 48)	0
Batch normalization (Batch Normalization)	(None, 48)	192
dropout (Dropout)	(None, 48)	0
dense (Dense)	(None, 24)	1176
dropout_1 (Dropout)	(None, 24)	0
dense_1 (Dense)	(None, 24)	600
dropout_2 (Dropout)	(None, 24)	0
dense_2 (Dense)	(None, 3)	75

Dataset Properties:

Our newly developed academic domain student feedback dataset were annotated for Roman Urdu sentiment analysis tasks and categorized into three groups, i.e., positive, negative, and neutral. Unfortunately, there is no publicly accessible dataset that can be used as a benchmark and contains student reviews with relative polarity values recorded in the ground truth. Consequently, dataset of 2000 student reviews was applied, which includes 1064 positive, 514 negative, and 422 as illustrated in Table 5. The dataset is distributed in an 80%:20% ratio into training and test set respectively. In the dataset, the shortest review has 4 to 5 words, and the most comprehensive review has 42 words. This distribution of reviews shows that most reviews are between 1 and 12 words.

Table 5. Student's feedback dataset, split evenly between a training set and test set

Sentiment class	Training set	Test set	∑
Positive (1)	925	139	1064
Negative (0)	391	123	514
Neutral (2)	284	138	422
∑	1600	400	2000

RUSA-19 Dataset:

The RUSA-19 dataset consists of 10,021 (Roman Urdu) customer reviews on a variety of subjects, including drama, technology, food, software, sports and blogs. The data were acquired from different social media platforms. RUSA-19 corpus includes 3302 neutral reviews, 3778 positive reviews and 2941 negative reviews.

Annotation Process & Guidelines:

In order to annotate the student reviews manually, the entire annotation procedure is discussed in this section. The procedure entails creating manual annotation rules, estimating inter-annotator agreement, and manual annotation. Three native speakers of Urdu who have a thorough understanding of Roman Urdu, especially, were chosen to annotate the dataset. Annotators were given a set of instructions to follow while annotating.

Guidelines for Positive Class:

If every aspect of the student review of the assigned text is positive, the review is considered positive. A student review is a collection of one or more sentences that express feelings toward a teacher. Positive class selected if the student reviews are mixed with a neutral and positive attitude. All reviews were marked as positive if they directly use the words "good," "great," "wonderful," "brilliant," or "fantastic." Positive words like “acha” (“good”), “madad-gar” (“helping”), “azeem” (“great”), “shandar” (“wonderful”) and “laajawab” (“outclass”) should be employed rather than negative ones like "Na," "Nahi," and "mat," as these terms flip the polarity[32].

Guidelines for Negative Class:

A sentence is considered negative if a certain word expresses a negative feeling or if there is a strong sense of apparent conflict. A text review is identified as negative when negative sentiment predominates over other sentiment categories. Sentences that include unpleasant, depressing, or sadness or include negative keywords such as “bura” (bad), “sakht”(harsh), “bad-ikhlaaq” (ill-mannered), “bakwass” (rubbish), “badtameez”(rude), without any additional modifiers like "Na," "Nahi," or "mat"; is perceived as negative[32]. All the unhappy, angry, or violent reviews are likewise labeled as negative.

Guidelines for Neutral Class:

Factual sentences, or those in which any idea is expressed, are considered neutral. Statements like "shayad" (Perhaps) and "ho saktaa hey" (Maybe) are examples of neutral statements since they convey a low degree of certainty and liability. A phrase expressing positive and negative opinions on specific characteristics and entities is said to be neutral[32].

Data Annotation Process:

A manual annotation process were followed in which the annotators manually add annotations to the student feedback dataset. Three annotators (x, y, and z) were chosen for tagging the dataset to create a new students' feedback dataset. Experts who have a solid understanding of Roman Urdu and Urdu have annotated the entire corpus. According to the provided annotation rules, each annotator has been given a review to categorize the review into three groups (positive, negative, or neutral). To evaluate the efficacy of the annotation procedure, a sample of 300 reviews annotated by X and Y were checked for contradicting pairs. In the context of a controversy, annotator Z had assigned to be a third annotator to assign a label to the controversial review. Additionally, an (IIA) inter-annotator agreement between annotators has been developed for the entire corpus.

Evaluation Metrics:

Precision, recall, F1-measure, and accuracy were utilized as evaluation measures to assess the performance of the suggested model. In order to increase the effectiveness of model in experimental setup, the fully connected layer and dropout layer were mainly focused with a specific size of neurons at each layer. It is observed that the maximum length of the tokens is 67 size used at neuron input size and applies a positive effect on the performance of the model rather than using neuron size in the sequence like 128,64, 32 and so on. RMSprop and ADAM are used as optimizer function with the BiLSTM model, ADAM is slightly better than

RMSprop with this dataset features. Finally, the categorical cross-entropy is used with the Softmax layer to predict close values towards three groups Positive, Negative and Neutral.

Findings and Discussion:

In this part, the findings of our experiments performed on two datasets of Roman Urdu are discussed, one is our newly created students feedback dataset other is RUSA-19[31] customer reviews dataset for sentiment analysis. Under two alternative neural word embedding techniques, our suggested model significantly beats the benchmark deep learning models.

Table 6. Performance of BiLSTM model using enhanced embedding

Dataset	Model	Embedding	Accuracy	Precision	Recall	F1-Score
Student's feedback	BiLSTM	Word2Vec	90.4	0.90	0.89	0.90
		GloVe	89.9	0.89	0.88	0.89
RUSA-19[31]	BiLSTM	Word2Vec	0.73	0.71	0.68	0.72
		GloVe	0.71	0.69	0.67	0.71

Comparison with Baseline Method:

In order to assess the model and draw conclusions about the findings, this section compare our model to some previously used baseline models in order to evaluate the effectiveness of the model.

Table 7. Results of the proposed model and baseline models with enhanced word embedding technique

Model	Word Embedding	Accuracy	Precision	Recall	F1-Score
CNN	Word2Vec	0.85	0.84	0.84	0.85
	GloVe	0.86	0.88	0.87	0.87
RNN	Word2Vec	0.82	0.84	0.83	0.84
	GloVe	0.81	0.72	0.72	0.72
GRU	Word2Vec	0.89	0.90	0.88	0.89
	GloVe	0.87	0.87	0.85	0.86
LSTM	Word2Vec	0.88	0.87	0.84	0.85
	GloVe	0.87	0.88	0.86	0.86
BiLSTM	Word2Vec	0.90	0.90	0.89	0.90
	GloVe	89.9	0.89	0.88	0.89

Two well-known feature representation techniques, word2vec and glove are used to assess the effectiveness of all models. Accuracy, recall, F1-score, and precision are the evaluation measures that are used. It is evident that CNN performed effectively, achieving F1-Scores of 0.85 and 0.87 for two distinct feature representation techniques. It is important to note that whereas CNN obtained a higher F1-score of 0.87, under the same representation classic RNN performed poorly for the GloVe feature representation, with F1-score of 0.72. On the other hand, GRU[3] performed better overall, achieving F1 scores of 0.89 and 0.86. Contrary to expectations, the traditional LSTM[21] has also demonstrated performance comparable to CNN, RNN, and GRU. The LSTM obtained F1-scores of 0.85 for word2vec and 0.86 for GloVe. Our suggested approach performed better than existing models and greatly enhanced performance. Two embedding techniques and all four evolutionary measures showed improved performance for our proposed model. Our recommended model, however, performed the best when using the word2vec embedding scheme, with precision scores of 0.90, recall scores of 0.89, F1-scores of 0.90, and accuracy scores of 0.90. Our model was evaluated on the publically available RUSA-19[31] Customer Reviews dataset to see if the suggested approach is generalizable. The classification results obtained from the model on this dataset, shown in Table 6, looks reasonable with an accuracy of 0.73 with word2vec embedding. The main reason in the performance difference of both data sets is colloquial nature of text, due to uneven and informal representation of Roman Urdu text, Stemming & normalization process discussed in the preprocessing section was performed on student's feedback dataset to reduce inconsistency and complexity of data. In case of RUSA-19 dataset, such type of stemmer hasn't been developed because of the huge amount of data. So, it is observed from the experiments that stemming process gives positive effect on the performance of the model.

Discussion

Deep learning techniques typically require more learnable time, data, and parameter tuning than machine learning techniques. At the same time, deep neural networks' capacity to semantically empower the model is made possible by pattern recognition skills. In addition, promising outcomes are achieved with the help of neural word embedding in such networks.

Pre-trained Word2Vec and GloVe embedding's were employed that learned on an ample amount of annotated data to enhance the quality of the word vector. It plays an important role in fine-tuning the word embedding size and embedding dimensions in creating the syntactic association between the attributes and, hence, improving the performance. In addition, compared to alternative methods of completing the Sentiment analysis task, word2vec, a neural word embedding, has demonstrated more incredible performance. Feature extraction and classifier design are two crucial components of sentiment analysis tasks. Although LSTM has proven highly effective in many NLP problems, it still has room for improvement. To analyze sentiments in Roman Urdu, a deep learning model is proposed, based on a bidirectional LSTM with an enhanced word embedding technique.

The performance of the traditional RNN was also evaluated, which suffers from memory limitations problem. If the sequence duration is large, the conventional RNN won't be able to glean the necessary context information from prior timestamps. So, the classification accuracy of classical RNNs is a little bit limited by their inability to deal with long-term temporal relationships. The gradient vanishing problem is a challenging scenario that LSTM handles more effectively. On the other hand, the CNN method of handling text features is unreliable since the features are chosen in an arbitrary sequence, restricting the system's ability to grasp the meaning of sentences. The GRU and LSTM (Long Short-Term Memory) networks are often considered conceptually equivalent. However, the LSTM is preferable because of the reduced number of gates in GRU. The traditional LSTM illustrates that memorizing the background information is not enough to understand the semantics of a sentence, so our model suggests a bi-directional strategy that is more appropriate to address this problem. As indicated, our proposed model can better preserve context for Roman Urdu evaluations.

Conclusion

It is very complex to perform sentiment analysis tasks on data that is of a colloquial type. The most critical issues here are feature extraction and classifier design. Another ability that helps the deep learning model in terms of semantics is pattern recognition, most standard deep learning models are not able to do this. The traditional resolved this issue at some extent, but still there is a need of improvement. So, a proposed BiLSTM model performs better with enhanced word embedding word2Vec using Roman Urdu corpus. The enhanced word embedding is implemented to give the model more semantic power and improve its capacity for extracting patterns. The next step in this investigation is to compare several deep network configurations (CNN, RNN, GRU and classic LSTM) under already trained neural word embedding to determine which deep learning technique is best for handling the sentiment analysis task for Roman Urdu. It is clearly percieved from the experiments that BiLSTM with word2Vec neural embedding performed slightly better compared to GloVe embedding as shown in Table 6

Since Roman Urdu is a very casual method of written communication, people often employ multiple spellings for the same terms, leading to a wide variety of word forms and a high degree of multilingualism. Some people make unnecessary spelling changes when attempting to convey their emotions through the written word. For example, for the sentence “Teacher is so good (in English),” in Roman Urdu, it is like “Teacher buht acha hy”; some people might write alternatively "Teacher boooooot achaaa hy." However, this is easily managed through a robust stemmer but Roman Urdu unfortunately lacks. As discussed in the preprocessing section, the stemming & normalization process based on the dictionary has been applied to reduce the complexity and anomalies of data and it is observed from the experiments that the model performance is improved due to the stemming process. Comparatively, in case of the RUSA-19 dataset, general pre-processing techniques were applied instead of applying manual stemmer and normalization. So, this is the main reason in the performance difference of both datasets.

At the end, Sentiment polarity is predicted using the final Softmax output layer; which is positive, negative, or neutral.

Limitations and Future Work:

The task of sentiment polarity identification was successfully performed well by our system. However, there were a few comments that were misclassified. The model will be improved by adding a fusion of another model with BiLSTM. Additionally, the current system only deals with comments given in the Roman Urdu. It was noticed that most of the students gives feedback in English and Roman Urdu, thus the system would be further expanded by processing Bilingual (English and Roman Urdu) reviews.

Acknowledgment

The Universiti Teknologi Malaysia (UTM), Malaysia, is sincerely acknowledged by the authors for its assistance.

Authors’ Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been
- included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in Universiti Teknologi Malaysia.

Authors’ Contribution Statement

N. designed and developed the Sentiment Analysis Model and collected the sample data. S. H. reviewed, proofread and critically analyzed the manuscript.

References

1. Qutab, I., K.I. Malik, and H. Arooj. Sentiment Analysis for Roman Urdu Text over Social Media, a Comparative Study. arXiv preprint arXiv.2020;16408.
2. Khan, I.U., et al., A review of Urdu sentiment analysis with multilingual perspective: A case of Urdu and roman Urdu language. Computers. 2021; 11(1): p. 3. https://doi.org/10.3390/computers11010003

3. Mehmood, F., et al., A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis. IEEE Access. 2020; 8: p. 192740-192759.https://doi.org/10.1109/ACCESS.2020.3030885
4. Masroor, H., et al., Transtech:development of a novel translator for Roman Urdu to English. Heliyon. 2019; 5(5): p. e01780. https://doi.org/10.1016/j.heliyon.2019.e01780
5. Poria, S., E. Cambria, and A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network. KBS. 2016; 108: p. 4249. http://dx.doi.org/10.1016/j.knosys.2016.06.009
6. AL-Bakri NF, Yonan JF, Sadiq AT. Tourism companies assessment via social media using sentiment analysis. Baghdad Sci. J. 2022 Apr 1; 19(2):0422. http://dx.doi.org/10.21123/bsj.2022.19.2.0422
7. Khan, M. and K. Malik. Sentiment classification of customer's reviews about automobiles in roman urdu. in Advances in Information and Communication Networks: Proceedings of the 2018 Future of Information and Communication Conference (FICC) .2019;Vol. 2. Springer.
8. AL-Jumaili AS. A hybrid method of linguistic and statistical features for Arabic sentiment analysis. Baghdad Sci. J. 2020 Mar 18; 17(1 (Suppl.)):0385. https://dx.doi.org/10.21123/bsj.2020.17.1(Suppl.).0385
9. Wang, W., et al. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In Proceedings of the AAAI Conference on Artificial Intelligence. 2017. https://doi.org/10.1609/aaai.v31i1.10974
10. Liao, S.N., et al., A robust machine learning technique to predict low-performing students. ACM transactions on computing education (TOCE). 2019; 19(3): p. 1-19. https://doi.org/10.1145/3277569
11. Chauhan, G.S., P. Agrawal, and Y.K. Meena, Aspect-based sentiment analysis of students' feedback to improve teaching–learning process. ICT. 2019; Springer. p. 259-266. https://doi.org/10.1007/978-981-13-1747-7_25
12. Ali, F., et al., Transportation sentiment analysis using word embedding and ontology-based topic modeling. KBS. 2019; 174: p. 27-42. https://doi.org/10.1016/j.knosys.2019.02.033
13. Atzeni, M. and D.R. Recupero, Multi-domain sentiment analysis with mimicked and polarized word embeddings for human–robot interaction. Future Gener. Comput. Syst. . 2020; 110: p. 984-999. https://doi.org/10.1016/j.future.2019.10.012
14. Dessí, D., et al., Deep learning adaptation with word embeddings for sentiment analysis on online course reviews. in deep learning-based approaches for sentiment analysis. 2020; Springer. p. 57-83. https://doi.org/10.1007/978-981-15-1216-2_3
15. Kaibi, I., E.H. Nfaoui, and H. Satori, Sentiment analysis approach based on combination of word embedding techniques. in Embedded Systems and Artificial Intelligence. 2020; Springer. p. 805-813.https://doi.org/10.1007/978-981-15-0947-6_76
16. Birjali, M., M. Kasri, and A. Beni-Hssane, A comprehensive survey on sentiment analysis: Approaches, challenges and trends. KBS. 2021; 226: p. 107134. https://doi.org/10.1016/j.knosys.2021.107134
17. Yadav, A. and D.K. Sentiment analysis using deep learning architectures: a review. Artif. Intell. Rev. . 2020; 53(6): p. 4335-4385. https://doi.org/10.1007/s10462-019-09794-5
18. Rao, G., et al., LSTM with sentence representations for document-level sentiment classification. Neurocomputing. 2018; 308: p. 49-57. https://doi.org/10.1016/j.neucom.2018.04.04
19. Cheng, Y., et al., Text sentiment orientation analysis based on multi-channel CNN and bidirectional GRU with attention mechanism. IEEE Access. 2020; 8: p. 134964134975. https://doi.org/10.1109/ACCESS.2020.3005823
20. Abid, F., et al., Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter. Future Gener. Comput. Syst.2019; 95: p. 292-308. https://doi.org/10.1016/j.future.2018.12.018
21. Ghulam, H., et al., Deep learning-based sentiment analysis for roman urdu text. Procedia Comput. Sci. . 2019; 147: p. 131-135. https://doi.org/10.1016/j.procs.2019.01.202
22. Guo, S., et al., Improved SMOTE algorithm to deal with imbalanced activity classes in smart homes. Neural Process. Lett. . 2019;50(2): p. 1503-1526. https://doi.org/10.1007/s11063-018-9940-3
23. Chandio, B., et al., Sentiment Analysis of Roman Urdu on E-Commerce Reviews Using Machine Learning. CMES-Comput. Model. Eng. Sci, 2022. https://doi.org/10.32604/cmes.2022.019535
24. Kamyab, M., G. Liu, and M. Adjeisah, Attention-based CNN and Bi-LSTM model based on TF-IDF and glove word embedding for sentiment analysis. Appl. Sci. . 2021; 11(23): p. 11255. https://doi.org/10.3390/app112311255
25. Xu, G., et al., Sentiment analysis of comment texts based on BiLSTM. Ieee Access. 2019; 7: p. 51522-51532. https://doi.org/10.1109/ACCESS.2019.2909919
26. Chandio, B.A., et al., Attention-based RU-BiLSTM sentiment analysis model for roman Urdu. Appl. Sci. . 2022; 12(7): p. 3641. https://doi.org/10.3390/app12073641
27. Khan, L., et al., Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media. Appl. Sci. . 2022; 12(5): p. 2694. https://doi.org/10.3390/app12052694

28. Chandio, B., et al., Sentiment Analysis of Roman Urdu on E-Commerce Reviews Using Machine Learning. CMES-Comput. Model. Eng. Sci. 2022. https://doi.org/10.32604/cmes.2022.019535

29. Mahmood, Z., et al., Deep sentiments in roman urdu text using recurrent convolutional neural network model. Information Processing & Management.2020; 57(4): p. 102233. https://doi.org/10.1016/j.ipm.2020.102233.

30. Uysal, A.K. and S. Gunal, The impact of preprocessing on text classification. Inf. Process. Manage. .2014; 50(1): p. 104-112. https://doi.org/10.1016/j.ipm.2013.08.006

31. Khan L, Amjad A, Afaq KM, Chang HT. Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media. Appl. Sci. . 2022 Mar 4;12(5):2694. https://doi.org/10.3390/app12052694

32. Mehmood, K., Essam, D., Shafi, K., & Malik, M. K.. Sentiment analysis for a resource poor language—Roman Urdu. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). 2019;19(1): 1-15. https://doi.org/10.1145/3329709

تحليل المشاعر لدى طلاب اللغة الأردية الرومانية ردود الفعل باستخدام تقنية تضمين الكلمات المحسنة

نورين [1]، شارين هازلين حسبي[1]، ظفر علي[3،2]

[1]قسم الحوسبة التطبيقية والذكاء الاصطناعي، الجامعة التكنولوجية الماليزية، جوهور، ماليزيا.
[2]كلية رزاق للتكنولوجيا والمعلوماتية، الجامعة التكنولوجية الماليزية، كوالالمبور، ماليزيا.
[3]قسم علوم الحاسب الآلي، معهد إدارة الاعمال، جامعة سوكور، سوكور، باكستان.

الخلاصة

تعد تعليقات الطلاب أمرًا بالغ الأهمية للمؤسسات التعليمية لتقييم أداء معلميها، حيث يتم التعبير عن معظم الآراء بلغتهم الأم، خاصة للأشخاص في مناطق جنوب آسيا. في باكستان، يستخدم الناس اللغة الأردية الرومانية للتعبير عن آرائهم، وينطبق هذا في مجال التعليم حيث يستخدم الطلاب اللغة الأردية الرومانية للتعبير عن تعليقاتهم. إنها عملية تستغرق وقتًا طويلاً وتتطلب عمالة مكثفة للتعامل مع الآراء النوعية يدويًا. بالإضافة إلى ذلك، قد يكون من الصعب تحديد دلالات الجملة في نص مكتوب بأسلوب عامي مثل اللغة الأردية الرومانية. تقترح هذه الدراسة تقنية تضمين الكلمات المحسنة وتبحث في تضمين الكلمات العصبية (Word2Vec وGlove) لتحديد أيهما أفضل أداءً لتحليل المشاعر الرومانية الأردية. يستخدم نموذجنا المقترح شبكة BiLSTM للحفاظ على السياق في كلا الاتجاهين وفي النهاية، يتم الحصول على نتائج التصنيف الثلاثي باستخدام طبقة إخراج softmax النهائية. تم استخدام مجموعة بيانات تم تصنيفها يدويًا لتقييم النموذج، وتم جمع البيانات من مؤسسات التعليم العالي في باكستان. تم تقييم النموذج تجريبيًا على مجموعتي بيانات باللغة الأردية الرومانية، ومجموعة بيانات تعليقات الطلاب المطورة حديثًا ومجموعة بيانات RUSA-19 المتاحة للجمهور باللغة الأردية الرومانية. يعمل نموذجنا بفعالية باستخدام تضمين الكلمات وطبقة BiLSTM. تمت مقارنة النموذج المقترح مع النماذج الأساسية لـ CNN، RNN، GRU وLSTM الكلاسيكية. توضح النتائج التجريبية فعالية النموذج المقترح بدرجة F1 تبلغ 90%.

الكلمات المفتاحية: شبكة الذاكرة طويلة المدى ،الأردية الرومانية، تحليل المشاعر، تعليقات الطلاب، تضمين الكلمات.