SENTIMENT ANALYSIS OF STUDENT FEEDBACKS USING DEEP LEARNING METHODS

Submitted by:
Soroor Monzavi

Email :Soroor.Monzavi77@gmail.com

Date :11/30/2022

Submitted to department of computer engineering
In partial fulfillment of the requirements of the degree of bachelor of science

At the

BAHAI' INSTITUTE FOR HIGHER EDUCATION (BIHE)

Supervisor:
Mr. Vargha Dadvar

Dedicated to Mona Naqhib, the 8-year-old Baluch girl who was shot dead on her way back from school. Whose eyes reminded me that sometimes, feeling neutral is a sin.

Contents

Table of Figures 4

Table of Tables 4

AbstracT 5

Introduction: 6

Literature review: 8

English sentiment analysis: 8

Persian sentiment analysis: 9

Students feedback sentiment analysis: 10

Methodology: 11

Dataset: 11

Preprocessing: 13

a) Cleaning and Merging Files: 13

b) Splitting feedback to sentence: 13

c) Combination of Question and Answers: 14

d) Text Cleaning: 14

Machine Learning Models: 15

a) Introduction: 15

b) Baseline Model: 15

c) Simple LSTM: 15

d) Concatenated LSTM: 16

e) Glove: 17

Training and Evaluation: 17

a) Dataset split: 17

b) Imbalanced dataset: 18

c) Evaluation metrics: 18

Experiments: 19

Simple LSTM experiment results: 19

Concatenated LSTM experiment results: 22

Glove LSTM experiment results: 24

Result analysis and model comparison: 25

Conclusion: 27

Bibliography 28

TABLE OF FIGURES

Figure 1- Dispersion of the imbalanced dataset 12
Figure 2 - LSTM architecture resource : colah.github.io 16
Figure 3 - Two branch LSTM architecture 17
Figure 4 - Precision Formula 18
Figure 5 - Recall Formula 19
Figure 6 - Simple LSTM loss graph 21
Figure 7 - Simple LSTM accuracy graph 21
Figure 8 - Simple LSTM confusion matrix 22
Figure 9 - Two branch LSTM loss graph 23
Figure 10 - Two branch LSTM accuracy graph 23
Figure 11 - Two branch LSTM confusion matrix 24
Figure 12 - Glove LSTM loss graph 24
Figure 13 - Glove LSTM accuracy graph 25
Figure 14 - Glove LSTM confusion matrix 25
Figure 15 - Binary LSTM loss graph 26
Figure 16 - Binary LSTM accuracy graph 26

TABLE OF TABLES

Table 1 - Sample labeled input 7
Table 2 - Raw data fields name and data type 11
Table 3 - List of questions in the dataset 12
Table 4 - Neutral data sample 13
Table 5 - Separate neutral data sample 13
Table 6 - Dependency of feedback sentiment to question 14
Table 7 - Concatenated input, combination of question and feedbacks 14
Table 8 - Separate input, input1: question, input2: feedback 14
Table 9 - Test, train and validation size for each label 18
Table 10 - Test, train and validation class weights 18
Table 11 - Simple LSTM experiment results 20
Table 12 - Simple LSTM all metrics results 20
Table 13 - Two branch LSTM Architecture 22

ABSTRACT

Growth of textual comments and opinions for products, Institutional services, political or social events and actions creates a large set of unstructured data that can be used to predict public's opinion. Sentiment analysis is one of the tasks of NLP (mostly NLU) that aims to solve the problem of understanding the sentiment on text automatically with programmed methods.

Students feedback sentiment analysis is one of the important tasks of NLP that has received less attention than the other tasks (like product reviews or movie reviews) in sentiment analysis, with the growth of online learning institutions in the past years, and the growing amount of online forums that students participate in them to express their feelings and send their feedbacks, student feedback sentiment analysis can play an important role in the improvement of educational institutions.

This paper aims to implement and compare different variations of LSTM model for Student Feedback sentiment analysis with small datasets to find out what procedure will be useful.

INTRODUCTION:

Sentiment analysis as a classification problem can be solved using different approaches. The Machine learning approach has been one of the most popular ones in the past few years, with the improvement in deep learning and neural networks models have reached better performance and predictions. Sentiment analysis in general simplifies mining sentiments out of textual data by automating it. It has become famous for business owners or producers, because it eases product appearance analysis and provides real-time analysis, saves time with automation and removes human bias through consistent analysis. [1], [2]

Student feedback analysis and course evaluations have become one of the responsibilities of educational institutions, and as online universities become more popular (specially over covid-16 quarantine [3]) developing a system for evaluating these surveys has become important to improve the teaching methods or teaching content. Also student feedback informs teachers on the effectiveness of their practice and identifies areas for future professional learning [4] and enables them to better understand the students perspective [5].

The task of student feedback sentiment analysis has received less attention than product review sentiment analysis. And the data provided in this task has its own characteristics, textual feedback provided with comment box questions gives students an opportunity to express their real opinion more freely as Likert-scale questions don't cover them [6] . Because of the nature of the relationship between students and teachers, students have specific lexicon, They have their own linguistic allusions and have a large variety of specialized words, they might not strongly express their opinion and hide it or say it more gently [5] due to the existence of the university hierarchy as compared to product reviews.

Compared to English, the Persian language includes many different complexities compared to English. These challenges are lack of tools, data resources, wide variety of declensional suffixes, word spacing and many informal or colloquial words [7] The amount of cleaned, structured data, tools and lexicons to the task of sentiment analysis in English is more compared to Persian language [8]

Many English researchers had the amount of data needed to use neural network based approaches like CNN or LSTMs. Persian sentiment analysis has used more traditional approaches like naive Bayes and SVM or rule based approaches and not many Persian sentiment analyses have focused on neural network approaches.
The paper aims to do a comprehensive implementation for student feedback sentiment analysis in Persian language with deep learning approaches focused on LSTM models and compare the achieved accuracy to find a suitable implementation with acceptable results.

We faced many challenges during its research, one of the anticipated challenges was the lack of data, the raw data contains about 400 feedbacks that should be labeled in 3 classes. Another challenge arose in the pre-processing and preparing data to label them, each raw feedback is a

paragraph of combination of 2 or 3 sentences and each of the sentences has an independent sentiment, which makes the labeling non uniform and makes it difficult to train the model.

another challenge was the dependency of questions and answers for labeling, the content of the questions had direct impact on the sentiment of the answers. Table 1 shows a sample of the data that the sentiment of the feedback depends on the question.

Question	Feedbacks	Labels
چه نقاط ضعف مهم دیگری را در ارائه این درس مشاهده می کنید که جای اصلاح دارد؟	پروژه محور بودن درس	negative
چه نقاط قوت مشخص دیگری در ارائه این درس وجود دارد که خوب است حفظ و تقویت شود؟	پروژه محور بودن درس	positive

TABLE 1 - SAMPLE LABELED INPUT

Due to the unbalanced weight of different labels in the dataset, the model was weak in detecting and predicting neutral labels. as the number of negative and positive labeled data is 4 times the number of neutral labeled data.

There are two question that this paper wants to find answers for them:

- Can usual methods used for sentiment analysis and in general NLP tasks be useful for Persian student feedback sentiment analysis considering Persian language special characteristics? How are features and complexities of the Persian language going to affect the proposed method and results?
- What are the solutions for the lack of data issues, related labeling issues, poor performance of the models in neutral label detection and dependency of the question and answer for labeling in student sentiment analysis? What is the best?

LITERATURE REVIEW:

ENGLISH SENTIMENT ANALYSIS:

Sentiment analysis is a one of the popular tasks of NLP for the English language, many researches, datasets and tools are available for English sentiment analysis, and there are sufficient data sources for the English language.

With the growth of deep learning and its tools and applications in recent years most of the sentiment analysis systems proposed in English have used deep learning-based methods.

Proposed methods have used different word embedding techniques [9] [10] different architectures for neural networks, [11] converting word to vectors, [12] [13] [14]and improving the loss function of neural network model [15] to achieve higher accuracies.

Deep learning based approaches have shown that they can outperform state-of-art machine learning approaches [16]. most popular deep learning models were LSTM combined with CNNs RNNs and BLSTMs that were compared with classic machine learning models mostly Naive Bayes and SVMs. Most of the researches achieved accuracies given by deep learning approaches are higher (up to 90%) compared to SOA machine learning approaches (70%) [12] [17]

khuong vo et al. [15] combined the domain knowledge to improve efficiency and accuracy, two of the major problems raised in sentiment analysis in lexicon approaches and neural networks, are (i) the existing works have not paid attention to the importance of different types of sentiment terms (ii) the loss function currently employed does not well reflect the degree of error of sentiment misclassification. Khoung researches on their paper introduces two improvements using quadratic programming to learn sentiment score for data augmentation and using weighted cross entropy with penalty matrix as an enhanced loss function.

khoung's approach for using quadratic programming then using sentiment scores to perform augmentation of the dataset to train deep learning model can be a useful point of view for sentiment analysis in Persian sentiment analysis for students' feedbacks special task.

Qianzi shen et al. [12] does multiple experiment on simple LSTM and simple convolutional neural network (CNN) and compares the result an as a final solution combined CNN and Bidirectional Long Short-Term Memory (BLSTM) as a complex model to analyze the sentiment orientation of text, over a dataset of 50000 movie reviews and reached accuracy of 89.7%. The pre-processing in this dataset turned each word into a 50D word vector that improved the result compared to word embedding techniques.

LSTM is one of the powerful tools for text understanding and the combination of LSTM and CNN might be an effective method for sentiment analysis on students' feedback.

Georgios k et al. [16] focused on hate speech detection and the approach they employed was a neural network solution composed of multiple LSTM based classifiers and utilizes user behavioral characteristics such as the tendency towards racism or sexism to boost

performance. In this paper the hate speech detection task highlights the need to explore the user features more systematically to further improve the classification accuracy.

PERSIAN SENTIMENT ANALYSIS:

Compared to English sentiment analysis, sentiment analysis issues have been less discussed in the Persian language and the body of knowledge has not been sufficiently developed, lack of data resources and tools for Persian language has reduced the diversity of knowledge produced in this field. [18]

Most Persian sentiment analysis have used classic machine learning approaches [18] [19] [20] [21] or lexicon based approaches [7] [22] [23] due to lack of data. [24]

In lexicon based approaches, have used limited lexicons, tools (like SPerSent, CNRC, POS, SentiPers) and models like Naive Bayes, SVMs and K-nearest neighbors.

The presented approach in Amiri et al. [24] paper is designing and implementing a lexicon-based sentiment analysis method for Persian text, their experiments yield between 60-69% accuracy rates for the initial version of the lexicon based Persian sentiment analysis. The value in this approach in particular is that an acceptable accurate lexicon based approach can be used to bootstrap an ML based system that does not require a large training set to start achieving results.

Basiri et al. [18] worked on a Persian sentiment analysis and used a well-known machine learning method, naive Bayes to evaluate his data. they believed that a common problem in previous studies on sentiment analysis in Persian is the lack of comprehensive dataset, another common problem in existing datasets is that almost all of them provide positive/negative labels this limits researchers only to polarity detection tasks. so he presents a new sentence-level dataset for sentiment analysis in Persian, SPerSent, a large (around 150000 sentences) dataset a precise lexicon CNTC containing about 2500 sentiment terms and a new stop-word list precisely collected from the web that provides both polarities and strength labels for each sentence that solves two both addressed problems.

The weakness of Basiri research was that the method used in this paper, Naive Bayes, has been replaced by neural networks in English language research.

Ebrahimi et al. [7] claimed that in recent years' Persian sentiment analysis task had two approaches to deal with: dictionary base and corpus base, and the dictionary based approach has been less investigated, so the paper introduced a supervised and automated method for creating a sentiment dictionary in Persian language using a corpus based approach consisting of preprocessing Persian reviews and extracting linguistic features, part of speech tagging extracting subjectivity words, determining the semantic orientation of words and calculating score for sentiment words.

Nasrin Taghizadeh et al. [25]represented Bert model based on ParsBert which is trained over general Persian language data, fine-tuned over large Persian medical corpora. The

model has been fine-tuned over and evaluated on question classification, sentiment analysis and question retrieval tasks. The fine-tuned Bert outperformed the state-of-art Persian language models. The question retrieval task had better results than sentiment analysis task. However, Taghizadeh claims that the supervision that exists in the classification tasks somewhat closes the gap between their Bert and the other language models.

STUDENTS FEEDBACK SENTIMENT ANALYSIS:

The task of students' feedback sentiment analysis, also is another field that has received less attention compared to customer product reviews sentiment analysis. Because of the nature of the task and its special difficulties of data gathering, research on non-English languages is limited. Again classic machine learning approaches [25] [26]and rule based approaches (VADER) [25] [27]are more used than neural network approaches. [28] [29]

One of the challenges in the task of student feedback sentiment analysis is gathering structured or labeled data for machine learning models.

There is little research in students' feedback sentiment analysis that has compared neural network approaches with classic machine learning approaches.

Nasim et al. [5] describes a sentiment analysis model trained using TF-IDF and lexicon based features to analyze the sentiments expressed by students in their textual feedback. In contrast to lexicon-based approaches, the proposed approach of combining the use of sentiment lexicon with machine learning techniques was capable of predicting the sentiment of the textual content even if the opinion words do not exist in the lexicon.

This paper is important because of the similarity of the task but this paper is only focused on predicting the polarity of the sentiment and is a binary classification problem.

Ali et al. also worked on the task of sentiment analysis of students' feedback, the proposed model implemented on Multinomial Naive Bayes, Stochastic Gradient Descent, Support Vector Machine, Random Forest and Multilayer Perceptron Classifier approaches and the MNB outperform the rest with 83.30%, also this paper is focused on binary classification and have not considered the neutral class. Another difference is that Ali's paper is using classic machine learning approaches like most of the research on non-English papers.

METHODOLOGY:

DATASET:

The dataset used in this article is an official survey of the students of the computer engineering field of the BIHE university collected during 4 consecutive semesters which was given permission to access them by the council. Feedbacks are made anonymously and the name of the course for which the feedback was registered has been deleted. The questions are related to the quality of course presentation and the response of the teaching teams, the volume of assignments and the strengths and weaknesses of the courses and they require descriptive answers.

The dataset consists of questions and textual explanatory answers by real BIHE students and the date when the feedback is registered. The raw data consists of 406 unprocessed feedbacks answered 6 questions within 3 files.

Table 2 is the list of columns descriptions:

column	descriptions
ID	Unique integer for each row of feedback
Start-date, end-date	Date representing the duration of the student filling the survey
email	Email of the student who filled the survey , all set to “anonymous”
name	Name of the student who filled the survey, all set to null
Feedback	The real content of the feedback answered to a specific question. String in Persian language.

TABLE 2 - RAW DATA FIELDS NAME AND DATA TYPE

The data is labeled in 3 classes (negative-0, neutral-1 and positive-2) manually to be used for training, validating and testing models.

The data was labeled manually as

- negative - 0 for feedback that represents a negative emotional charge based on the question.
- Positive - 2 for feedback that represents a positive emotional charge based on the question

- Neutral - 1 for feedback that does not express a positive or negative feeling. (either It is a suggestion or it has shared information, and in some cases represents both positive and negative feelings in such a way that it is impossible to separate them in the sentence.)

After the cleaning and preprocessing that data, the proportion of the labels in the original data are shown in figure 1 :

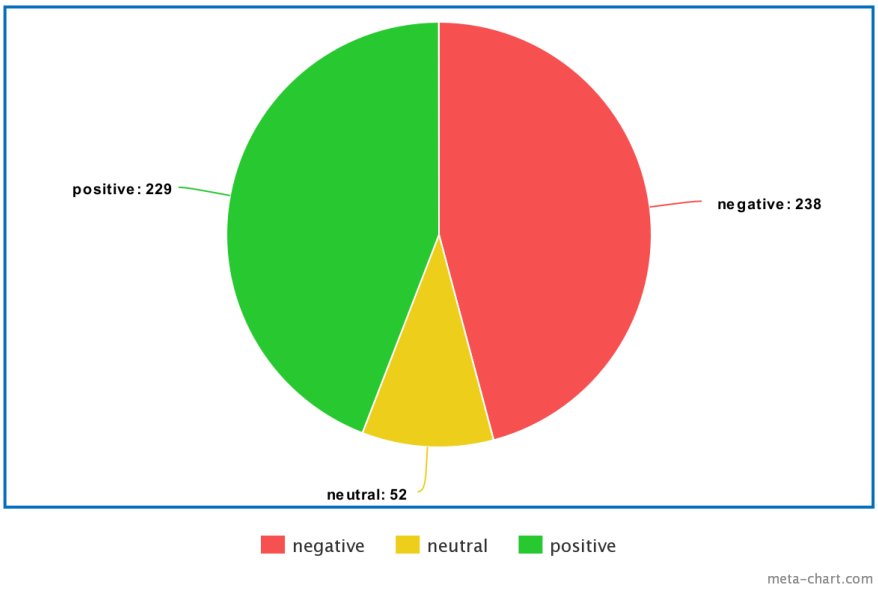

FIGURE 1- DISPERSION OF THE IMBALANCED DATASET

Feedbacks are answers to total 6 questions that are showing in table 3:

Index:	Questions:
1	آیا سرعت و کیفیت پاسخگویی تیم آموزشی به سؤالات و نیازهای دانشجویان و همین طور ارائه بازخورد به تکالیف و فعالیت ها مناسب بوده است؟
2	از بین منابع و فعالیت های آموزشی درس، مانند کلاسهای آنلاین، ویدئوهای آموزشی، تکالیف، امتحانات، کلاس های حل تمرین، کارگاه، فروم ها و گروه های درسی، اسلاید های آموزشی و کدام موارد بیشترین تاثیر را دارد ؟
3	چه نقاط ضعف مهم دیگری را در ارائه این درس مشاهده می کنید که جای اصلاح دارد؟
4	چه نقاط قوت مشخص دیگری در ارائه این درس وجود دارد که خوب است حفظ و تقویت شود؟
5	حجم و کیفیت تکالیف و فعالیت های درس را چطور ارزیابی می کنید؟
6	کلاس ها و منابع آموزشی درس تا چه حد در راستای یادگیری های شما کافی و مفید بوده است؟

TABLE 3 - LIST OF QUESTIONS IN THE DATASET

PREPROCESSING:

a) CLEANING AND MERGING FILES:

The cleaning process of the data started from removing extra columns, like ID, start date, end date, email, name.

Then all 3 files were merged inside another file, with 2 columns for “question” and “answer”, for each remaining column title (the question) and every row of each column (the answer to that question) of each file, a row was added to the merged file.

b) SPLITTING FEEDBACK TO SENTENCE:

Due to the nature of the data, each feedback contained several sentences with different sentiments. Assigning a positive or negative sentiment to each feedback was inconclusive. In order for the labels to be chosen more definitively and specifically, each feedback of the raw data is split by the “.” to several sentences and after that labeled according to new data.

After a secondary review of data and models results, feedbacks with neutral labels were also split by “اما” and “ولی” to represent separate sentences with specific sentiments.

Sample:

One real feedback is in table 4:

Feedback	label
در نیمه ی دوم ترم استاد درس تغییر کرد. استاد جدید بسیار قابل و با معلومات بودند اما کمی در فعالیت های درس بی نظمی دیده میشد	Neutral - 1

TABLE 4 - NEUTRAL DATA SAMPLE

One sample After split is table 5:

feedback	Label
در نیمه ی دوم ترم استاد درس تغییر کرد.	Neutral - 1
استاد جدید بسیار قابل و با معلومات بودند	Positive - 2
کمی در فعالیت های درس بی نظمی دیده میشد	Negative - 0

TABLE 5 - SEPARATE NEUTRAL DATA SAMPLE

c) Combination of Question and Answers:

The sentiment of each feedback is related to the question that was asked shown in table 6.

Question	Feedback	Label
حجم و کیفیت تکالیف و فعالیت های درس را چطور ارزیابی می کنید؟	بسیار زیاد تر از حد انتظار بود	Negative - 0
آیا سرعت و کیفیت پاسخگویی تیم آموزشی به سؤالات و نیازهای دانشجویان و همین طور ارائه فیدبک به تکالیف و فعالیت ها مناسب بوده است؟	بسیار زیاد تر از حد انتظار بود	Positive - 2

Table 6 - Dependency of feedback sentiment to question

According to this condition, the models were fed with two types of data.
One with concatenated strings of question and answer into the simple model to analyze the sentiment for the combination of question and answer shown in table 7.

Input	حجم و کیفیت تکالیف و فعالیت های درس را چطور ارزیابی می کنید؟بسیار زیاد تر از حد انتظار بود

Table 7 - Concatenated input, combination of question and feedbacks

Another with two separate inputs of question and answer into concatenated model with two LSTMs analysis the sentiment for question and answer separately and then get an overall sentiment for the combination of two labels shown in table 8.

Input 1	حجم و کیفیت تکالیف و فعالیت های درس را چطور ارزیابی می کنید؟
Input 2	بسیار زیاد تر از حد انتظار بود

Table 8 - Separate input, input1: question, input2: feedback

d) Text Cleaning:

The data was cleaned from punctuations, null and duplicate values, also numeric values, single characters and English string were removed to manipulate data.
The data was normalized and lemmatized with the help of the Hazm library.
Stop words were used from the short list of stop words of Kharazi/Persian stop words. [30]
Since the list of stop words in most of the Persian library contains suffixes and prefixes that removing them from the real feedback sentence changes the sentiment of verbs from their actual label the stop word list was checked and unacceptable values were removed from it.

MACHINE LEARNING MODELS:

a) INTRODUCTION:

The focus of this paper is on testing different architectures of LSTM models for the task of sentiment analysis for Persian students pre-processed feedback in 3 classes.
We proposed a long short term memory (LSTM) based recurrent neural network given that RNN approaches outperforms the state of art auto regressive based models [31].
RNNs are powerful tools for modeling sequence data, RNN models remember its input, due to an internal memory, which makes it perfectly suited for machine learning problems that involve sequential data. To solve the problem of vanishing gradients we used LSTMS. [32]

b) BASELINE MODEL:

The Logistic regression was used for the baseline model, as a supervised learning model, a linear model used when the dependent variable is categorical. Logistic regression is also simple and efficient and fast for first experiments. due to the smallness of the dataset and to avoid complex models with many parameters that results overfit.
a simple Logistic regression for multi classifications was implemented and was fed with concatenated question and feedback to receive the simplest result with **0.607** score.

c) SIMPLE LSTM:

LSTMs provide us with a large range of parameters such as learning rates, and input and output biases. Hence, no need for fine adjustments. The complexity to update each weight is reduced to O(1) with LSTMs. similar to that of back propagation through time (BPTT), which is an advantage.

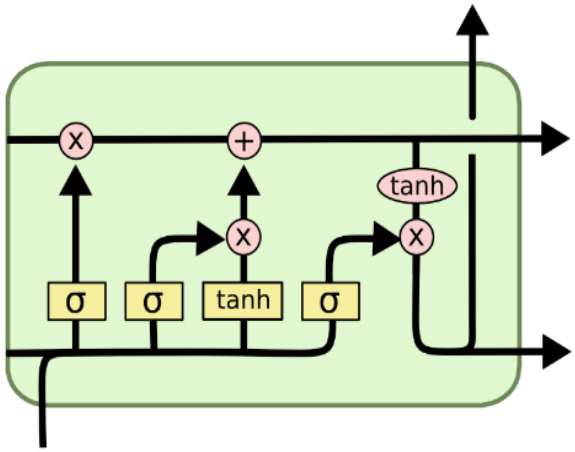

FIGURE 2 - LSTM ARCHITECTURE RESOURCE : COLAH.GITHUB.IO

The simplest version of the LSTM implemented is a sequential model containing an embedding layer, a simple lstm-64 layer and an activation layer with the activation function set to softmax. Other versions have different values of embedding parameters, dropouts, more dense layers and more LSTM layers with different parameter values.
this LSTM receives a sequence of inputs and returns a single character representing the predicted label.
The embedding layer designed in these sequential models receives input as a sequence of words in question concat with the feedback.
The models were compiled with the two different callbacks, stopping early to avoid overfit and checkpoints to save the best model over large batches.

d) CONCATENATED LSTM:

One of the challenges described in the introduction section is the dependency of questions and answers sentiment. The solution in Simple LSTM was to look at the input as a string containing question and answer and predicting an overall sentiment of the input.

To design a more precise neural network that can predict the sentiment of question and answers in their own context, a two-branch model is designed that learns question and answers separately and predicts a sentiment with respect to two different inputs.

The concatenated LSTM implemented in this paper is a neural network containing two branches of LSTMs for question and answer each, with two different embedding input layers receiving tokenized question and tokenized answer separately. Each branch has a LSTM with a different input size. the output of two LSTM branches are merged into one dense layer. the activation layer set to softmax. Figure 3 is an illustration of the implemented model.

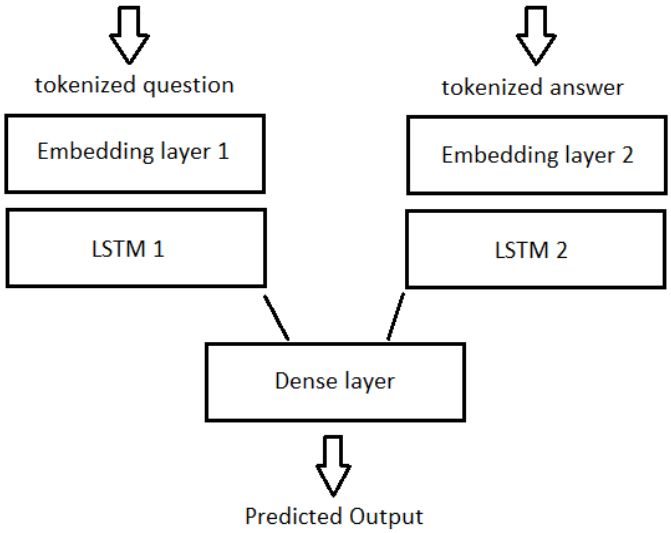

FIGURE 3 - TWO BRANCH LSTM ARCHITECTURE

e) PRE TRAINED MODEL:

due to the lack of data challenge, and after testing simple and concatenated data, the glove layer was tested on one simple LSTM as a pre-trained model.

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. [33]

The glove vector, is added as a weight matrix to a non-trainable embedding layer.

This pre-trained embedding layer increases the speed of learning, and helps the model to save data on learning the data specific characteristics and the general characteristics are included in the glove layer.

We have tried to test pre-trained models on sentence level embedding and compare the results for any improvements but finding a reliable and well-designed sentence level embedding layer in Persian didn't have much results.

TRAINING AND EVALUATION:

a) DATASET SPLIT:

In order to avoid indeterminacy of the result, the original and prepared data were divided into 3 (train-validate-test) datasets. With the rate of 25% of all data for test dataset and 25% of the rest of the data for validation and the rest for training. All 3 datasets were divided with respect to labels so that all sub datasets have the same weights for 3 labels following in table 9.

dataset	Number of Negative	Number of Neutral	Number of Positive
Train	133	29	129
Test	60	13	57
Validation	45	10	43

TABLE 9 - TEST, TRAIN AND VALIDATION SIZE FOR EACH LABEL

With the same class weights following in table 10.

Dataset	Negative class weight	Neutral class weight	Positive class weight
Train	0.729	3.344	0.751
Test	0.722	3.333	0.760
Validation	0.725	3.266	0.759

TABLE 10 - TEST, TRAIN AND VALIDATION CLASS WEIGHTS

b) IMBALANCED DATASET:

due to the fact that the neutral labels in the dataset are less than positive and negative labels, the model has less sensitivity to detect neutral labels and the prediction is not reliable. The high accuracy in these cases usually shows the bias of the model on negative and positive labels.

to avoid the model bias due to imbalance data, models were fit with class weights to balance the results according to each class weights.

c) EVALUATION METRICS:

Used metrics for analyzing the performance of models predicting sentiments for feedback, were precision and recall. which both suit multi-class classification predictions perfectly with the OVR (one vs rest) strategy. [34]

- precision answers the question of "what proportion of predicted positives are truly positive?"

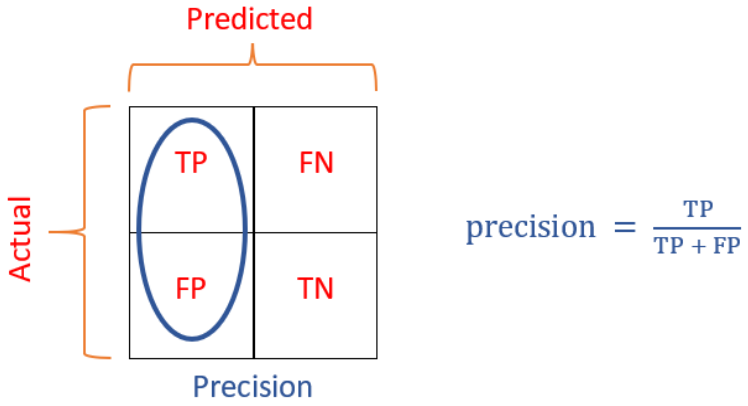

FIGURE 4 - PRECISION FORMULA

- recall answers the question of "what proportion of the actual positives are correctly classified?"

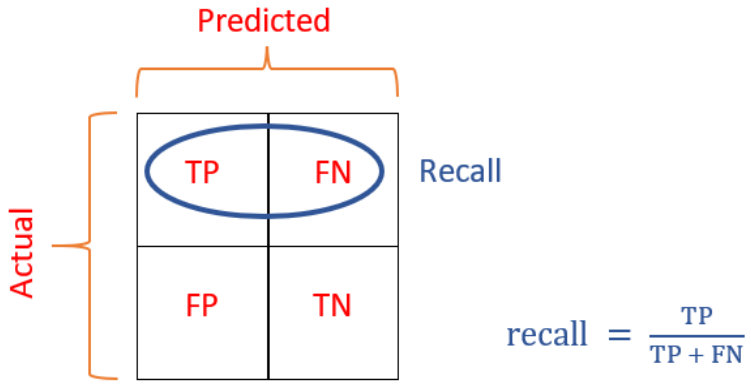

FIGURE 5 - RECALL FORMULA

EXPERIMENTS:

The data fed to these models to experiment, were the same resource explained in section Methodology.2 in two ways, one with concatenated question and answers, used in 1.simple LSTM and 3.glove LSTM, another with separate questions and answers used in 2.concatenated LSTM.

All models used ModelCheckPoint callbacks to save best accuracy among epochs, also the fit method used class weight to help model balance the imbalance dataset.

SIMPLE LSTM EXPERIMENT RESULTS:

First LSTMs were designed as simple as possible and with new observations, changes were made and they became more complete and precise.

Table 11 is the loss and accuracy results on test result of different simple LSTM.

classifier	max-num-words	output dim	LSTM size	drop out	dense	loss	accuracy
classifier 1	332	128	64	0.4	32	0.667	0.677
classifier 2	332	64	64	0.4	32	0.749	0.7
classifier	332	64	128	0.4	32	0.508	0.677

3							
classifier 4	332	64	128	-	32	0.677	0.654
classifier 5	332	64	128	0.4	64-32	0.506	0.692
classifier 6	322	64	128	0.4	-	0.492	0.685

TABLE 11 - SIMPLE LSTM EXPERIMENT RESULTS

the classification report table for classifier 5 is as shown in table 12:

classifier 5	precision	recall	f1-score
negative	0.76	0.58	0.66
neutral	0.21	0.23	0.22
positive	0.67	0.82	0.74

TABLE 12 - SIMPLE LSTM ALL METRICS RESULTS

This table shows that the version 5 of simple LSTM with increasing the LSTM units to 128 and decreasing output units to 64 with two dense layers with 64 and 32 units followed by two 0.4 drop outs, was able to get the best results.

This table shows the simple LSTM model does not perform well for neutral cases and for negative and positive cases it can do better. The data loss and imbalance data is still a challenge in simple LSTM and the model has bias in prediction neutral sentiments.

The following shows the accuracy and loss plot for simple LSTM that is trained over 50 epochs that leads to over fit and the ModelCheckPoint saves the best weights for best accuracy in epoch 48.

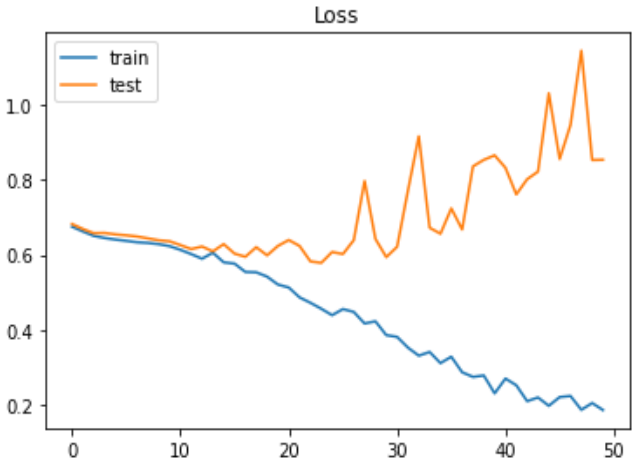

FIGURE 6 - SIMPLE LSTM LOSS GRAPH

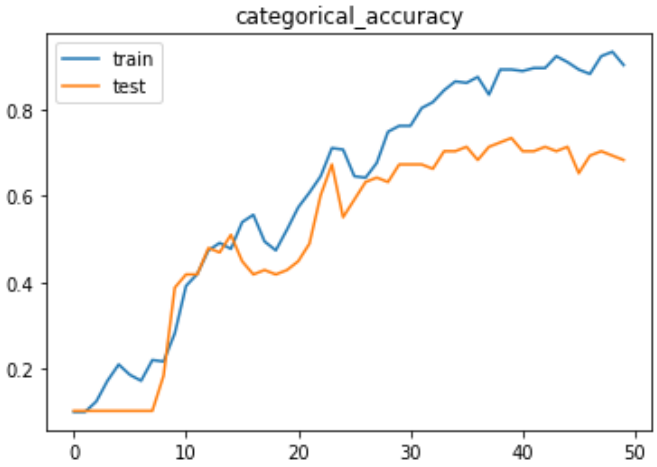

FIGURE 7 - SIMPLE LSTM ACCURACY GRAPH

The results of the confusion matrix show that the model predicts positive or negative more often than neutral; this problem is consistent with the small amount of neutral data and the "class weight" parameter did not solve the whole imbalance dataset.

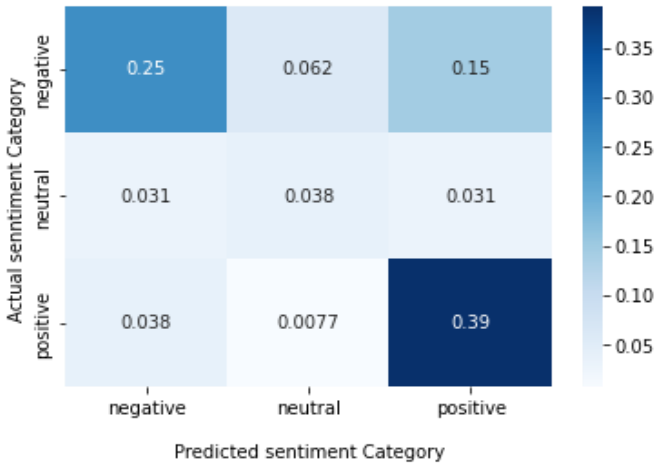

FIGURE 8 - SIMPLE LSTM CONFUSION MATRIX

CONCATENATED LSTM EXPERIMENT RESULTS:

The concatenated model was implemented by creating two branches with two separate input layers, designed one for question inputs and another for answer inputs. each with one LSTM size 64 and then they concatenated with one dense 64 following the attention layer.

This is the architecture implemented in this paper shown in table 13:

input 1	input 2	output
embedding (386)	embedding (190)	drop out (0.4)
LSTM (64)	LSTM (64)	dense (64)

TABLE 13 - TWO BRANCH LSTM ARCHITECTURE

The LSTM reached 0.698 accuracy and 0.654 as the best results on two branch LSTM.

The following shows the accuracy and loss plot for two branch LSTM that is trained over 50 epochs that leads to over fit and the ModelCheckPoint saves the best weights for best accuracy at epochs 42.

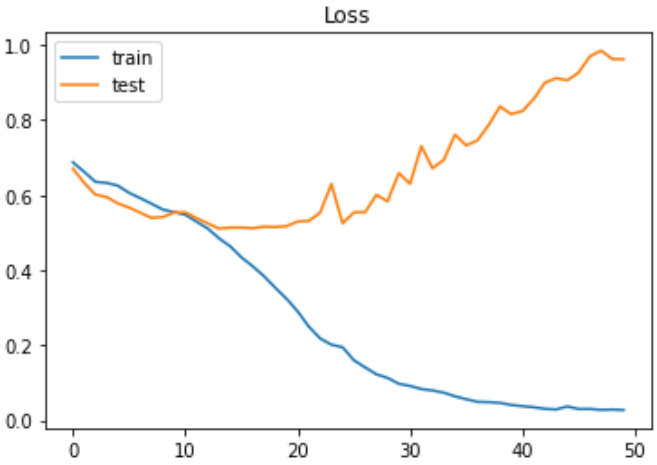

FIGURE 9 - TWO BRANCH LSTM LOSS GRAPH

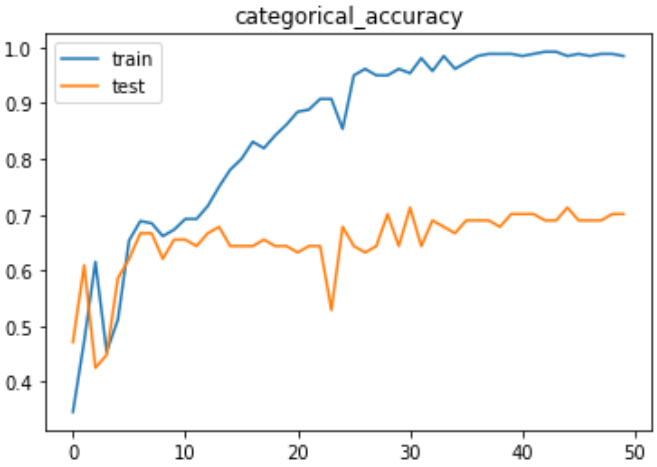

FIGURE 10 - TWO BRANCH LSTM ACCURACY GRAPH

The result of two branch LSTM confusion matrix shows that the model predicts positive or negative more often than neutral. Also the “class weight” solution did not solve the problem for two branch LSTM as like the simple LSTM.

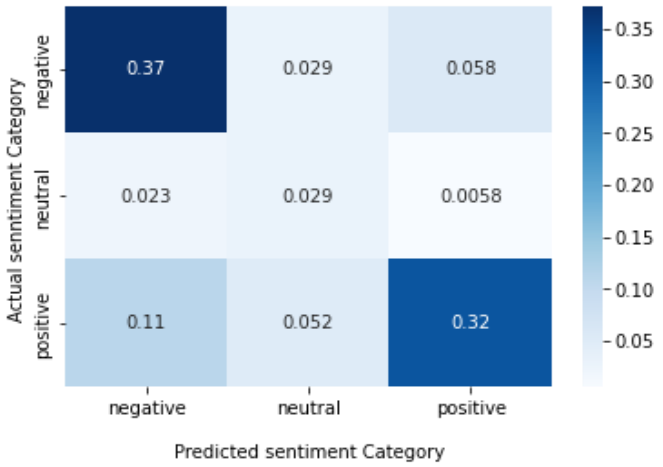

FIGURE 11 - TWO BRANCH LSTM CONFUSION MATRIX

GLOVE LSTM EXPERIMENT RESULTS:

The glove layer uses a pre-trained embedding matrix that eases the process of training for the model, in this case it should improve the performance of the model, so that the model will be pre-trained by the embedding matrix and the actual data should fine tune the model.

The LSTM reached 0.563 accuracy and 0. 600as the best results on Glove LSTM.

The following table shows the loss and accuracy of the glove model. The glove model improves the loss and accuracy from first epochs because of the pre-trained layer in the embedding layer.

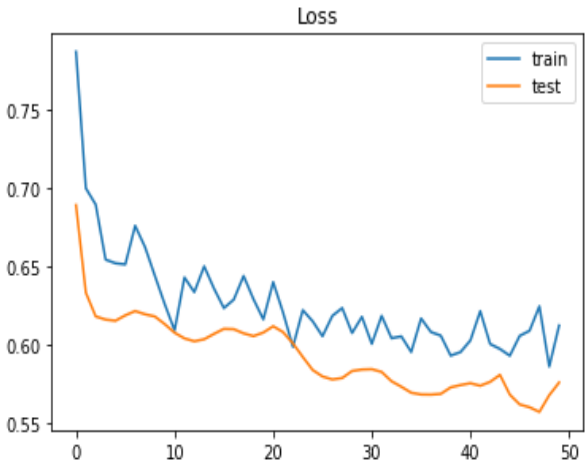

FIGURE 12 - GLOVE LSTM LOSS GRAPH

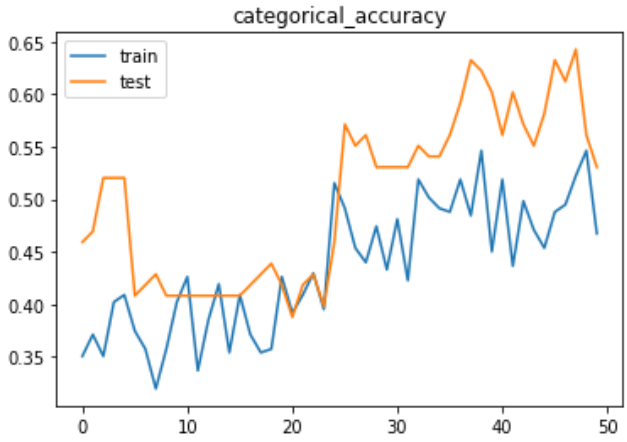

FIGURE 13 - GLOVE LSTM ACCURACY GRAPH

The results glove model LSTM confusion matrix shows that the model predicts positive or negative more often than neutral. Also the "class weight" solution did not solve the problem for glove LSTM as like the simple LSTM.

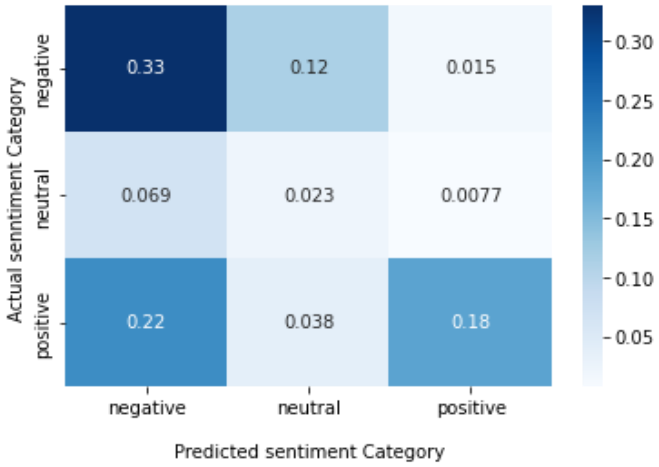

FIGURE 14 - GLOVE LSTM CONFUSION MATRIX

RESULT ANALYSIS AND MODEL COMPARISON:

The concatenated LSTM and glove LSTM were expected to have higher performance for data with two inputs and data small datasets.

The concatenate model was predicted to ease the process of learning through using more learning units for different inputs and leave each LSTM branch with its special data to help the model be more precise in learning. but the results were not improved with two branch LSTM.

The glove model was supposed to speed the progress of learning by using the pre-trained embedding matrix, to save the students feedback for fine tuning the model with the actual data characteristics. The glove was used to reach better performance with less data.

The loss plot from the glove model was expected to start the loss value from a lower value compared to other models with no pre-trained layer, but not much change and improvement was found in the train and test plots. The plots start from improved values but the slope of changes is low and the final value has not been changed compared to previous results.

one possible reason might be special characteristics of the final data that did not match pre trained embedding layer characteristics. The results show that there was not much progress.

as the results show that the model acts poorly for prediction neutral sentiments, in another experiment the model was called on dataset where neutral labels were removed and model was implemented for a binary classification the results improved:

The LSTM reached 0. 348accuracy and 0. 812as the best results on Binary LSTM.

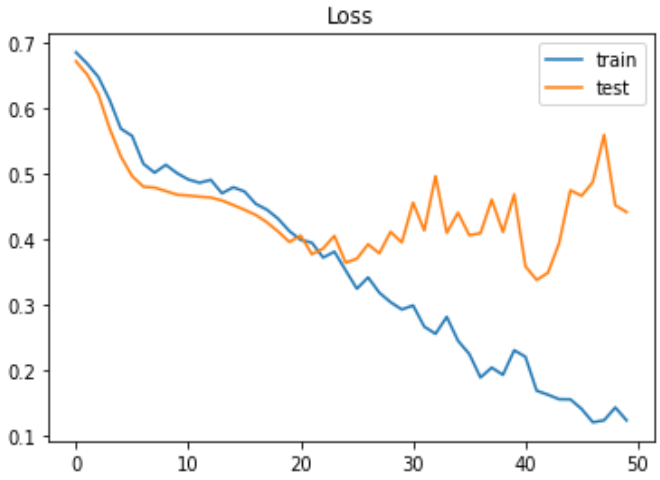

FIGURE 15 - BINARY LSTM LOSS GRAPH

FIGURE 16 - BINARY LSTM ACCURACY GRAPH

CONCLUSION:

sentiment analysis has been used in many industries and helped institutions and businesses to improve their services and products with monitoring a large number of customers' opinions and feedback.

universities are one of the institutions that needs to be constantly criticized and modified based on the opinion of its audience. Due to the popularity of online education for students and universities, students have a lot of space to comment on forums and express their feelings and opinions.
The aim of this paper is to use a neural network based approach to implement a sentiment analysis to predict positivity , negativity and neutrality of the students feedback of BIHE through several semesters.

This paper used LSTM as the base model to use for implementation and used other variations of LSTM with two branch LSTM for two questions and feedback as the two input branches, and the glove LSTM to help the process of learning with small amounts of data.

The simple LSTM and two branch LSTM have small differences and improvements and the glove did not improve the performance as it was predicted.

The Bert implementation for sentiment analysis might be a good future work field to test, it was out of the context of time and focus of this paper.

Student sentiment analysis on Persian datasets still can be a good topic for future works, with focus on collecting data for Persian student feedbacks, finding better and more precise ways to clear sentence, separate them to simple short sentence with specific sentiments. And solutions for better predicting the neutral sentiment.

BIBLIOGRAPHY

[1] B. Liu, Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, 2010.

[2] A. Dorfman, "5 Real-World Sentiment Analysis Use Cases," *Reputation,* 2021.

[3] F. L. Cathy Li, "The COVID-19 pandemic has changes education forever.This is how," *Word Economic Forum,* 2020.

[4] L. Mandouit, "Using student feedback to improve teaching," *Educational Action Research,* pp. 755-769, 2018.

[5] S. H. Q. R. Zarmeen Nasim, "Sentiment analysis of student feedback using machine learning and lexicon based approches," *ResearchGate,* 2017.

[6] A. Dobronte, "Likert Scales vs. Slider Scales in commercial market research," *CheckMarket,* 2012.

[7] N. A. Faranak Ebrahimi Rashed, "A Supervised Method for Constructing Sentiment Lexicon in Persian Language," *ResearchGate,* vol. 1, 2017.

[8] M. S. F. d. J. Ayoub Bagheri, "Sentiment Classification in Persian: Introducing a Mutual Information-based Method for Feature Selection," in *Iranian Conference on Electrical Engineering (ICEE)*, 2013.

[9] X. X. X. C. D. W. Y. L. Z. Y. Hui Du, "Asepct-Specific Sentimental Word Embedding for Sentiment Analysis of Online Reviews," in *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016.

[10] W. Z. e. al., "Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis," *IEEE Transaction on Knowledge and Data Engineering,* vol. 30, no. 1, pp. 185-197, 2017.

[11] A. M. Abdolraouf Hassan, "Deep Learning approach for sentiment analysis of short texts," in *3th International Conference on control, Automation and Robotics (ICCAR)*, Nagoya, Japan, 2017.

[12] Z. W. Y. S. Qianzi Shen, "Sentiment Analysis of Movie Reviews Based on CNN-BLSTM," in *IFIP Advances in Information and Communication Technology*, Springer, Cham, 2017, pp. 164-171.

[13] T. I. S. W. S. G. Chetanpal Singh, "A Deep Learning Approch for Sentiment Analisys of COVID-19 Reviews," *Artificial Intelligence Computing and Applications for COVID - 19,* vol. 12, no. 8, 2022.

[14] P. K. S. Rani, "Deep Learning Based Sentiment Analisys Using Convolution Neural Network," *Arabian Journal for Science and Engineering,* vol. 44, pp. 3305-3314, 2018.

[15] D. P. M. N. T. M. T. Q. khoung Vo, "Combination of Domain Knowledge and Deep Learning for Sentiment Analysis," *Arxiv,* 2019.

[16] H. R. H. L. Georgios K. Pitsilis, "Detecting Offensive Language in tweets Using Deep Learning," *Arxiv,* 2018.

[17] J. W. R. Y. K. S. Z. C. Shiyan Liao, "CNN for sentiment understanding based on sentiment analysis o twitter data," *Procedia Computer Science,* vol. 111, pp. 376-381, 2017.

[18] A. k. Ehsan Basiri, "Sentence level Sentimnt Analysis in Persian," *ResearchGate,* 2017.

[19] A. B. MH Saraee, "Feature selection methods in Persian sentiment analysis," *Lecture Notes in Computer Science,* 2013.

[20] C. G.-C. D. Z. E Vaziripour, "Analyzing the political Sentiment of Tweets in Farsi," *ICWSM,* 2016.

[21] M. G. A. A. A. H. A. A. T. D. Kia Dashtipour, "A Comparative Study of Persian Sentiment Analysis based on different Feature Combinations," *Lecture Notes in Electrical Engineering,* vol. 463, 2018.

[22] A. R. N.-N. Ehsan Basiri, "A Framework for Sentiment Analysis in Persian," *Open Transactions on Information Processing,* pp. 1-14, 2014.

[23] A. A. R. H. M. M. A. S. A. M. Pedram Hosseini, "SentiPers: A Sentiment Analysis Corpus for Persian," *Arxiv,* 2018.

[24] S. S. M. H. K. Fatemeh Amiri, "Lexicon-based Sentiment Analysis for Persian Text," *RANLP,* 2015.

[25] K. K. M. A. C. H. N. L. Irfan Ali, "Student Feedback Sentiment Analysis Model Using Various Machine Learning Schemes A Review," *Indian Journal of Science and Technology,* 2019.

[26] M. M. G. M. C. Nabeela Altrabsheh, "SA-E: Sentiment Analysis for Education," in *The 5th KES International Conference on Intelligent Decision Technologies*, Portugal, 2013.

[27] R. L. Marion Neumann, "Capturing Student Feedback and Emotion in Large Computing Courses: A Sentiment Analysis Approch," in *52nd ACM Technical Symposium on Computer Science Education*, Virtual Event-USA, 2021.

[28] I. U. S. S. F. M. k. A. H. Muhammad Zubair asghar, "Fuzzy-Based Sentiment Analysis system for Analyzing student Feedback and Satisaction," in *preprints*, 2019.

[29] D. P. K. Sangeetha, "Sentiment Analysis of Student Feedback using Multi-Head Attention fusion Model of Word and Context Embedding for LSTM," *Journal of Ambient Intelligence and Humanized Computing,* 2021.

[30] kharazi.

[31] W. Z. X. T. M. Z. J. S. V. l. Z. L. J. Z. Xin Huang, "LSTM Based Sentiment Analysis for Cryptocurrency Prediction," vol. 12683, pp. 617-621, 2021.

[32] N. Donges, "A Guide to Recurrent Neural Networks: Understanding RNN and LSTM Networks," 2022.

[33] R. S. C. D. M. Jeffrey Pennington, "Glove: Global Vectors for Word Representation," *nlp stanford,* 2015.

[34] BEXGBoost, "Comprehensive Guide to MultiClass Classification Metrics," *Towards Data Science,* 2021.