IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 4, December 2024, pp. 4747~4756
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i4.pp4747-4756

Sentiment analysis of student's comments using long short-term memory with multi-head attention

Bhavana P. Bhagat[1], Sheetal S. Dhande-Dandge[2], Sandeep R. Raut[3]
[1]Department of Computer Engineering, Government Polytechnic, Yavatmal, India
[2]Department of Computer Science and Engineering, Sipna College of Engineering and Technology, Amravati, India
[3]IB Computer Science, Shanghai Qibao Dwight High School, Shanghai, China

Article Info

Article history:

Received Feb 11, 2024
Revised Apr 9, 2024
Accepted Jun 14, 2024

Keywords:

Convolutional neural network
Deep learning
Learning long short-term memory
Opinion mining
Sentiment analysis

ABSTRACT

Classroom teaching is a viable and effective approach for enhancing student learning and promoting engagement in the educational process. The opinions of students play a vital role in the evaluation of teachers. This paper presents a comprehensive overview of sentiment analysis techniques based on recent research and subsequently explores machine learning, i.e., ensemble classifiers, deep learning, long short-term memory (LSTM), convolutional neural network (CNN), LSTM with single attention, LSTM with multi-head attention, and feature extraction techniques (TFidfVector and Word2Vec), in the context of sentiment analysis over student opinion datasets, i.e., the Vietnamese student feedback corpus, as well as data collected from a final-year student's comment in 2023. Further, the Vietnamese student feedback corpus is translated to English and pre-processed with the proposed framework, which yields interesting facts about the capabilities and deficiencies of different methods. In this paper, we conducted experiments with ensemble classifiers, LSTM and CNN, LSTM with single attention, and LSTM with multi-head attention. We conclude that LSTM with multi-head attention produces an accuracy result of 95.57%, which outperform as compare to other three baseline methods and earlier studies.

Corresponding Author:

Bhavana P. Bhagat
Department of Computer Engineering, Government Polytechnic Yavatmal
Dhamangaon Road, Pimpalgaon, Yavatmal, Maharashtra 445001, India
Email: bhavana.bhagat30475@gmail.com

1. INTRODUCTION

Classroom teaching is a viable and efficient approach for enhancing student learning and promoting engagement in the educational process. The main objective of teaching is to improve students learning abilities and promote effective knowledge in them. Teaching effectiveness always depends on teachers' subject knowledge and pedagogical approach to teaching. Sangeetha and Prabha [1] stated that the quality of teaching relies on pedagogical skills, defined objectives, clarity, and subject expertise. Evaluating teaching effectiveness can be measured by collecting student feedback, aiding in the refinement of strategies for improving teaching excellence. Student's opinions are always deep and unique. They give an in-depth view of how their teachers encourage and educate. The teachers get opportunities to feel and value the importance of teaching from student's opinions. Student feedback helps the teachers formulate better decisions on how to improve the quality of teaching. Thus, students are one of the most important participants in every educational institute.

Researchers have developed a number of methods and tools to measure student opinions. In this era, sentiment analysis is one of an emerging area in the fields of natural language processing (NLP) and artificial intelligence. It involves the use of computational methods to determine and identify the sentiments or emotional attitudes expressed in text, either in student's feedback, news articles, social media posts, customer reviews, and other forms of written comments. It is used as a tool to analyze and evaluate users' comments automatically. According to Zhou and Ye [2], sentiment analysis has become one of the most significant topics in education research, and there are several implications and research issues discussed in the research related to the use of sentiment analysis in education. This study addresses the use of machine learning and deep learning technologies in education to classify sentiments from student comments using four different classification models and compares their performance with existing studies. Furthermore, this study aims to encourage researchers to focus on deep learning approaches to improve the performance of classifier models for sentiment classification.

Section 2 will briefly explain related work. Section 3 begins with an outline of the detailed methodology of the four approaches used in our proposed work. Section 4 presents experimental results and performance evaluations and comparisons with competing methods. Section 5 will discuss the conclusion and future scope.

2. RELATED WORK

In this work, we focus on sentiment analysis of student comments in the education domain. Some studies [3], [5]-[8] used a lexicon-based approach; some studies [5], [6], [8]–[10] did not evaluate results using evaluation metrics. Also, there is a trend to use hybrid models to achieve better results in the sentiment analysis process. Some models do not provide graphical results for sentiment analysis, nor do they provide any facts or studies to evaluate their visual usability. Many researchers in [11]–[15] focus on e-learning. Limited studies use the primary corpus, machine learning and deep learning methods, along with evaluation metrics [16].

Several researchers in [4], [9], [10], [17]–[25] employed machine learning-based methodologies. The model is initially trained with known inputs and outputs, in order to function with later unknown data. As stated in Bhagat and Dandge [16], some algorithms, perform more accurately than others when used with large corpora. In some studies [21], [26]–[29] hybrid methods were used to improve the accuracy of their work. Lastly, it is seen that several recent studies [1], [12], [30]–[34] used deep learning models as one of the dominant models in the education domain with higher performance. Many studies used the student feedback data, which is available online to explore and is human-annotated in Vietnamese, as in Nguyen *et al.* [30].

The proposed system implements machine learning and deep learning models and compares their results, to resolve the problem of text classification and compare their accuracy. Also used to identify the best model based on evaluation metrics for student's comments. The comparative analysis states that the deep learning methods achieve better than the rest of the three baseline models. The proposed system uses two datasets for training and testing, i.e., the Vietnamese student feedback corpus and primary dataset collected from a final-year student's comment. The education field is experiencing a significant and profound change. Within this sector, sentiment analysis holds great significance for key stakeholders, such as students and educational institutions. A variety of machine learning and deep learning techniques will be employed to evaluate student's feedback.

According to Sangeetha and Prabha [1], deep learning is popular currently due to its automatic learning capabilities and also it deals with long sequences rather than short sequences of sentences. It is observed that in previous work, a Vietnamese student feedback corpus consisting of 16,175 feedback sentences was used with the long short-term memory (LSTM) with multi-head attention model whereas in our proposed models, both machine learning and deep learning models were implemented, and testing was performed on both the Vietnamese student feedback dataset as well as unseen data collected from final year student feedback. Also, comparative analysis of deep learning and machine learning methods is conducted in our experiments to identify the best methods in terms of accuracy performance. In this work, we first converted the Vietnamese dataset to English, containing 21,561 sentences and trained it using a machine learning and deep learning model. We utilize both the Vietnamese and the primary dataset, which contain 520 comments from final-year students gathered in 2023 for model testing.

3. METHODOLOGY

The models proposed in this study include several steps for sentiment analysis. First, preprocessing and various word embedding techniques were used to generate a vector representation. Then, these vectors are used for training different classifier models for sentiment analysis. Finally, the results obtained with the proposed model are compared with those obtained with the other three baseline models in this study and previous studies. Figure 1 illustrates the overall methodology of the proposed sentiment classification models, which use different word embedding techniques and evaluation metrics to measure the performance of the system.

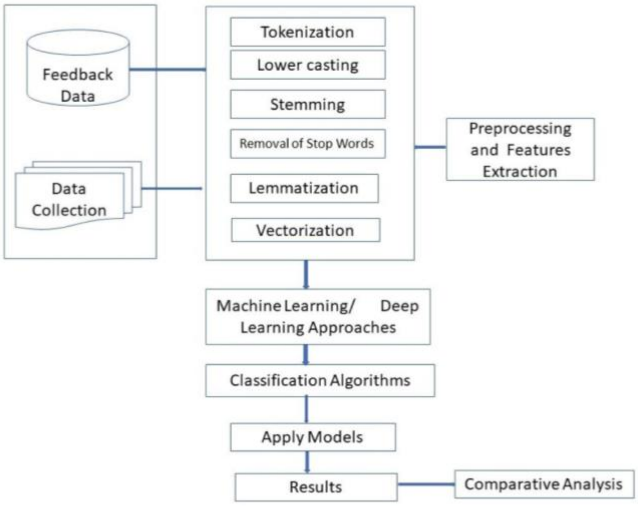


Figure 1. The methodology used for proposed models

3.1. Dataset

Evaluated model's performance on two data corpuses, the Vietnamese student feedback corpus by [30] and primary data collected on our own from a final-year student's comment. In our model, 21,561 rows of comments were used for sentiment analysis and 520 student's comments for the prediction of unseen data. We allocate 75% of each dataset class for training and 25% for testing.

3.2. Preprocessing and word embeddings

On this work, the student feedback dataset is used. It has some primal data, such as abbreviations, spelling mistakes, emotions, signs, repeated and unwanted characters, numbers, and symbols. To clean the text data, a text preprocessing method is used. Stemming is used for reducing a word to its stem, which affixes to prefixes and suffixes or to the roots of words called "lemmas". It is used to normalize the text and make it easier to process. Lemmatization analyzes the context and converts the word to its meaningful form, which is called a lemma. The bag of words model, which is called BoW, is used for extracting features from the text to be used in modeling. A tokenizer is used to divide paragraphs and sentences into smaller parts that can be easily assigned meaning.

Deep learning algorithms accept a vector representation of numbers as input. Word embedding is a technique for performing the representation of words in a vector space. It is a technique for translating words into a mathematical domain where numbers capture the semantics of the words. According to Pacol and Palaoag [4], the process is done using millions of learning and machine phrases. Word2vec is based on an unsupervised machine learning method based on a large amount of text to achieve syntactic and semantic representation in [35]. It is used with the LSTM and convolutional neural network (CNN) models, and the most common encoding method such as TFIDFvectorizer, has been used with the rest of our three proposed models.

3.3. Classification techniques

In this proposed work, four different sentiment classification approaches were used to classify the sentiments of students from student's comments. Various word embeddings techniques were used for feature extraction. These models were built, including machine learning such as ensemble classifier and deep learning such as CNN and LSTM, LSTM with attention layer, and LSTM with multi-head attention layer.

3.3.1. Ensemble classifier

The first approach uses an ensemble classifier. It is a decision tree ensemble made out of a random selection of decision trees. It is a tree predictor combination in which each tree is dependent on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In this, a decision tree has been used as a basic learner. Each tree was constructed using bootstrap samples from the training data. A random feature selection provides variation among the base learners. Bagging is used to reduce the variation of algorithms with a high variance. A noise or an outlier in a single tree classifier can affect the classification model's performance; however, random forest merged various weak learners to build a strong learner that can handle noise or outliers due to the randomness it can give. It is also capable of handling large datasets with a large number of attributes. More crucially, it retains accuracy even when data is unavailable [36].

3.3.2. Convolutional neural network and long short-term memory

The second approach uses CNN and LSTM. CNNs are a special type of multilayer neural network that learns from data. CNNs are the same as neural networks, having weights, biases, and input with which they make a scalar product and apply an activation function. Due to the increased number of hidden layers, there is a need for more computational resources; hence, overfitting arises. Thus, CNNs help to overcome problems by dividing models into small parts of information and combining this information in the deepest layers of the neural network. They are very useful for identifying classes, objects, and groups in images by looking for shapes in them. This framework has convolutional, pooling, and completely connected layers. A deep learning model needs a lot of computational power and data to train. As a result, CNNs were strictly constrained to limited sectors and were unable to enter the machine learning space [37].

LSTM is one of the most commonly used recurrent neural network (RNN) networks capable of learning long-term dependencies in classification models. The correct result is achieved when LSTM is extended with other models, and it may serve as the new base for additional models. The vanishing gradient problem, which affects neural networks, makes it challenging to comprehend the parameters of the previous layer. A LSTM in combination with other models helps to solve this issue. In the LSTM architecture, as stated in [1], [30], [38] the memory cells store and get information over long dependencies, consist of three gates called the input, forget, and output gates. The input gate i_t signifies the new information in the cell using input $x_{t,}$ the previous hidden layer, recurrent weight, and bias. A forget gate f_t determines the extent to which the previous memory cell is forgotten. The information obtained from the current input x_t, is given by the output gate $u_{t.}$ The input gate i_t, chooses which information from output gate u_t to store in memory cell c_t with a condition forget gate f_t and finally the output o_t of the LSTM depends on o_t and h_t.

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}) \tag{1}$$

$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}) \tag{2}$$

$$U_t = tanh(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}) \tag{3}$$

$$C_t = i_t.u_t + f_t.c_{t-1} \tag{4}$$

$$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}) \tag{5}$$

3.3.3. Long short-term memory with attention layer

Another approach used is LSTM with a single attention layer. It combines the single attention layer with the LSTM. It is employed to bring attention to the long sequence's referral word. By focusing on important information that relates to the target value, the attention layer has useful information and enhances model performance. Although it concentrates on the desired value, it doesn't exclude any essential information. In contrast to the other approaches, it can handle word sequences of different lengths. The attention model emphasizes semantic elements while also capturing the significance of each context word. They suggested an attention-based LSTM that outperforms the prior model in terms of output.

3.3.4. Long short-term memory with multi-head attention layer

Finally, our state-of-art-approach to sentiment analysis is LSTM with multi-head attention layer. As stated by [1], attention mechanisms have recently proven to be beneficial in a variety of NLP applications. Vaswani *et al.* [39] stated that the attention mechanism successfully handles sequence data. In this work, to increase performance and accuracy, we used LSTM with a multi-head attention mechanism, which integrates multiple heads. It learns a variety of unique features. A modern architecture for neural machine translation is multi head self-attention. Keys, values, and queries with n dimensions are managed by many heads. An attention function takes a query Q, a set of keys and values <K, V> to obtain an output O. This technique is often called scaled dot-product attention. It is a set of multiple heads that together learn various depictions at every position in the sequence, as mentioned in [40]. Figure 2 illustrates architecture of proposed system using multi-head attention with LSTM for sentiment analysis.

3.4. Comparative analysis

The comparative analysis states that deep learning models outperforms as compare to rest of studies in earlier works. The model was trained and tested with four different machine learning and deep learning algorithms along with word embedding techniques. The results of two datasets were reported using the confusion matrix: accuracy, precision, recall, and F1 score.

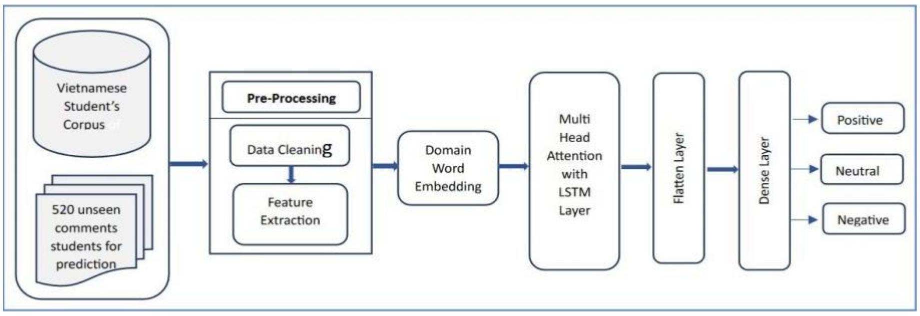


Figure 2. Architecture of proposed system using multi-head attention with LSTM for sentiment analysis

4. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

In order to find the best model, it is important to establish a series of steps that allow us to explore different model configurations, such as ensemble techniques, fully connected networks, dense and dropout layers, weight initialization and use of activation functions, learning rate with optimizers, as well as other hyper-parameters. It is also important to decide batch size, the number of epochs, and the evaluation metrics used. First, we pre-processed and tokenized the student's dataset, which contains positive, negative, and neutral sentiments, using feature extraction techniques (i.e., TFIDFVector and Word2Vec) to provide the tokenized form of the comments for our proposed approaches, which are given in Table 1. Word2Vec and TFIDFVectorizer were used to transform the tokanized comments into vectors. Finally, the models were trained using methods such as ensemble classifiers, LSTM and CNN, LSTM with single-head attention and LSTM with multi-head attention. Once the training and testing of student comments are completed, we compare our three bassline models to our proposed model, LSTM with multi-head attention and evaluate performance using accuracy, precision, recall, and F1-score.

Table 1. Number of sentiment labels in dataset

Sr. no	Sentiment	labels
0	lively	2
1	Although my test score is not high because I d...	1
2	Great work!	2
3	Teachers are very enthusiastic in teaching.	2
4	To spend time growing up to lecture for student	0

4.1. Ensemble classifier

In this approach, a hybrid classifier, which is a combination of multiple base classifiers using voting ensembles, is used. Five different base classifiers, each with a specific algorithm, were used. Firstly, two random forest classifiers with decision trees were used. Later, two AdaBoost classifiers with decision trees as weak learners were used, and finally, one gradient boosting classifier with decision trees was used. For multi-class classification problems, the 'OneVsRestClassifier' is used to perform multi-label classification, where each classifier is trained to handle a specific class, treating all other classes as a single "rest" class. The `VotingClassifier` is an ensemble method that combines the predictions of the base classifiers using majority voting ('soft' voting). The class labels with the highest combined probabilities among all base classifiers are selected as the final prediction.

4.2. Long short-term memory+convolutional neural network

In this approach, the model is evaluated using word embedding, pre-trained fine-grained Word2vec, LSTM and 1-D CNN. Two LSTM layers with 100 units each were added with a dropout of 0.3 in order to prevent overfitting. A 1D convolutional layer (`Conv1D`) with 100 filters and a filter size of 5 was added next, which performed 1D convolution operations on the input sequence. The activation function relu is used after the convolution layer. A 'GlobalMaxPool1D' layer takes the maximum value over the temporal dimension and reduces the dimensionality of the data, keeping the most important features. A dense layer with 16 units applies matrix multiplication to the input data and applies the relu activation function element-wise. A final dense layer and a SoftMax function are responsible for predicting the sentiment class probabilities.

4.3. Long short-term memory+single head attention

In this approach, the model is evaluated using LSTM with a single head attention layer. An LSTM layer with 64 units and return sequences will return the full sequence output rather than just the final output.

The attention layer helps the model focus on important parts of the input sequence during training. Next, a dropout layer with a dropout rate of 0.5 is added, which helps to prevent overfitting. Finally, a dense layer, i.e., an output layer with 3 units and a softmax activation function, is added, which is responsible for predicting the sentiment class probabilities. Also, the categorical crossentropy-based loss function, Adam as an optimizer, and an accuracy metric for evaluation are used in the model.

4.4. Long short-term memory+multi-head attention

In this proposed approach, the model is evaluated using LSTM with a multi-head attention layer. Figure 3 shows our experimental setup to classify sentiments of student's feedback into positive, neutral and negative using multihead attention with LSTM. An embedding layer is used to convert integer-encoded input data into dense vectors of a fixed size. It has 100 input dimensions (vocab size) and produces 20-dimensional embeddings. Next is a multi-head attention layer with heads equal to 3 and an embedding dimension equal to 128. A feed forward neural network (FFN) is implemented as a sequential model consisting of two dense layers. The first layer uses the ReLu activation function, and the second layer outputs a tensor with dimensions equal to 128. This FFN is used for further processing after multi-head attention. Next, layer normalization is applied to the outputs of the multi-head attention and feedforward layers. It helps to stabilize training by normalizing the activations within each layer. Dropout is a regularization method used during training to prevent overfitting, with a dropout rate of 0.5.

4.5. Evaluation metrics

According to Liu [41], the test indicators selected for our proposed models were accuracy, precision, recall, and F1-score. We performed experiments and compared the proposed LSTM with the multi-head attention model with other baseline models. The proposed LSTM with the multi-head attention classification model is compared with other baseline models as given: i) an ensemble classifier is a decision tree ensemble made out of a random selection of decision trees; ii) LSTM + CNN with more gates to control long sequences of information, along with word embedding in Word2vec; iii) LSTM+attention is a combination of LSTM and single attention layer used to focus the referral word in long sequences; and iv) LSTM with multi-head attention is a concatenation of embedding + multi-head attention + LSTM. It combines multiple heads to increase accuracy performance. It contains more layers than other base models. Two dropout and flatten layers were used to avoid overfitting.

4.6. Experimental results

In this section, Tables 2 to 5 display a confusion matrix and classification report for all four models. Table 6 shows the evaluation metrics, i.e., testing accuracy with reasonable performance accuracy, precision, recall, and F1-score for all models. It is seen that LSTM with multi-head attention performs better than the rest of the three models in terms of accuracy, F1-score, precision, and recall of 95.56%, 95.57%, 95.44% and 95.57% respectively, followed by LSTM+attention, ensemble classifier, and LSTM+CNN.

Table 2. Ensemble classifier confusion matrix

Actual	Predicted		
	Neutral	Negative	Positive
Neutral	1,523	26	38
Negative	0	1,761	337
Positive	4	51	1,651

Table 3. LSTM+CNN model confusion matrix

Actual	Predicted		
	Neutral	Negative	Positive
Neutral	1,788	75	145
Negative	228	1,621	207
Positive	205	125	2,075

Table 4. LSTM+single-head attention model confusion matrix

Actual	Predicted		
	Neutral	Negative	Positive
Neutral	1,694	8	156
Negative	91	1,431	71
Positive	85	18	1,837

Table 5. LSTM+ multi-head attention model confusion matrix

Actual	Predicted		
	Neutral	Negative	Positive
Neutral	5,218	5	83
Negative	224	4,589	245
Positive	127	33	5,646

Table 6. Evaluation metrics for all models

Metrices/models	LSTM + CNN	LSTM + ATT	Ensemble classifier	LSTM + multi-head attention
Accuracy (%)	84.778	92.04	87.18	95.56
F1-Score	0.8477	0.9204	0.9156	0.9557
Precision	0.8472	0.9190	0.94	0.9544
Recall	0.8477	0.9204	0.87	0.9557

Also, precision, recall, and the F1-score computed for all models were shown in Tables 7 to 9. It is seen that the performance on test data using precision for positive sentiments perform better in ensemble classifier and the performance on test data using recall and F1-score for neutral sentiments performs better in ensemble classifier as compared to the rest of three proposed models. We tested the model on unseen data collected from the 520 comments of final-year students enrolled in 2023 which also performed well as shown in Table 10.

Table 7. Performance on test data using Precision

Model	Precision		
	Neutral	Negative	Positive
LSTM+CNN	0.79	0.89	0.86
LSTM+ATT	0.91	0.90	0.95
Ensemble classifier	0.96	0.84	0.97

Table 8. Performance on test data using Recall

Model	Recall		
	Neutral	Negative	Positive
LSTM+CNN	0.84	0.85	0.86
LSTM+ATT	0.91	0.94	0.92
Ensemble classifier	0.98	0.89	0.88

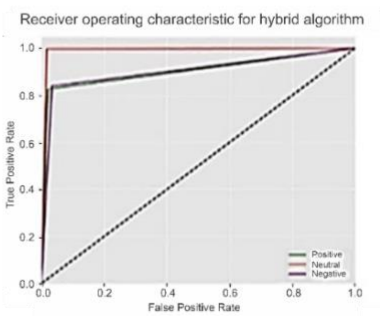
Table 9. Performance on test data using F1-Score

Model	F1-score		
	Neutral	Negative	Positive
LSTM+CNN	0.89	0.81	0.85
LSTM+ATT	0.91	0.98	0.89
Ensemble classifier forest	1	0.96	0.81

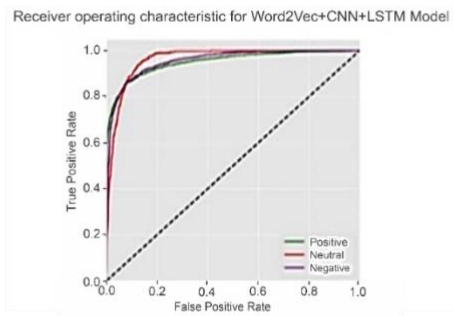
Table 10. Few sample predictions and actual values on unseen data

Input text	Actual sentiments	Predicted sentiments
Lecturers are easy to understand and enthusiastic	Positive	Positive
I feel pressured and uncomfortable to talk to you ask you but you don t answer	Negative	Negative
The teacher focuses on the right focus	Positive	Neutral
As for visual basic i don t think it s necessary so i shorten the teaching time instead of spending nearly half a semester to give an overview of visual basic	Negative	Negative
Apply images video clips in the teaching process	Positive	Negative
Sometimes a bit fast	Negative	Negative

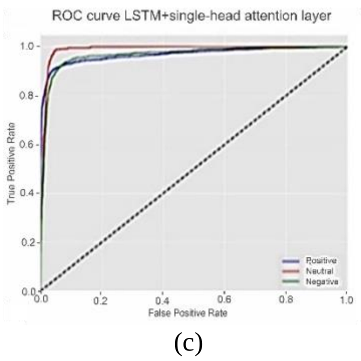
Figure 3(a) to 3(d) shows the receiver operating characteristic curve (ROC) curves for each class for all models, and the area under each curve (AUC) indicates the classifier's performance for that class. It shows the computation of the false positive rate, true positive rate, and threshold for each class. The function is applied to the true labels and predicted probabilities for each class and finally computes the area under the ROC curve for each class. A higher AUC value indicates that models are performing better compared to earlier works.



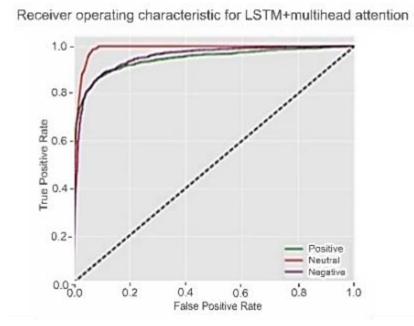
(a)



(b)



(c)



(d)

Figure 3. ROC curve for (a) ensemble classifier, (b) LSTM and CNN, (c) LSTM with single-head attention, and (d) LSTM with multi-head attention

4.7. Discussion and comparison models

Previous researchers used various lexicon-based, machine learning, and deep learning techniques to identify student's sentiments based on their responses. Also, some of the earlier studies did not provide evaluation metrics for performance measurement. These earlier works served as benchmarks against which we compared the results of our proposed models. In previous work by [1], [30] used the UIT-VSFC corpus and classification methods LSTM and DT-LSTM with SVM models with accuracy of 90.74% and FUSION models with accuracy of 91.96%, respectively, whereas our proposed model used LSTM with multi-head attention models with an accuracy of 95.75%. Also, our proposed model was tested on 520 unseen comments collected from the feedback of final-year students in 2023, which performed well on unseen data. Table 11 shows a comparison of our proposed model to the models from previous studies in which various deep learning approaches were used on student datasets as one of the dominant models used in the education domain for better performance. It demonstrates that our model LSTM with multi-head attention outperforms existing classifiers in earlier works as well as our proposed rest of the three benchmark methods.

Table 11. Comparison of our models to earlier studies using deep learning approaches

Study/year	Method/algorithm	Data corpus	Performance Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Sindhu *et al.* [32]/2016	LSTM model for layer 1 and layer 2	5,000 comments	93.0	88.0	85.0	86.0
Nguyen *et al.* [30]/ 2018	LSTM and DT-LSTM (with SVM) Models	UIT-VSFC corpus 16,175 sentences	90.74	NA	NA	90.2
Cabada *et al.* [31]/2018	CNN+LSTM	yelp -147672, SentiText - 10834, EduSere-4300 student's comments	SentiText- 88.26 and EduERAS - 90.30	NA	NA	NA
Sangeetha and Prabha [1]/ 2020	LSTM, LSTM+ATT, muitihead ATT and FUSION (multi-head ATT + Embedding + LSTM)	16,175 students feedback sentences	91.96	75.60	87.51	76.90
Mabunda *et al.* [27]/2021	Support vector machines, multinomial maive Bayes, random forests, k-nearest neighbours and neural networks	Students feedback from Kaggle dataset 185 records	84	NA	NA	NA
Our proposed models	Ensemble classifiers LSTM and CNN LSTM with attention LSTM with multi-head attention	The Vietnamese 21,561 student feedback corpus, as well as 520 student data collected from a final-year student's comment in 2023	87.18 84.77 92.04 **95.57**	94.0 84.0 91.0 **95.44**	87.0 84.0 92.0 **95.57**	91.0 84.0 92.0 **95.56**

*NA – Not available

5. CONCLUSION

We introduced the UIT-VSFC corpus and unseen final-year student's comments for the analysis of educational sentiment. We proposed the methods using a machine learning hybrid model, i.e., ensemble classifier, and deep learning models, i.e., LSTM and CNN models, along with word2vec embedding, LSTM with single attention, and LSTM with multi-head attention. According to the findings of our experiments, it was seen that our proposed model, LSTM with multi-head attention with embedding layer, achieves better results as per the experimental results conducted using other baseline methods as well as earlier models. Our model performs better for long sequences of sentences and also yields attention to the referral words. Moreover, by employing LSTM with a Multi-head attention layer, we achieve the most optimal results with testing accuracy, F1-score, precision and recall performance of 95.57%, 95.56%, 95.44%, and 95.57% respectively, which outperform the rest of the three baseline models and also produce better results in terms of accuracy, precision, recall, and F1-score as proposed by earlier studies. We also tested the model on unseen data collected from the 520 comments of final-year students. The model performs well on unseen data. In the future, we are going to perform experiments using variants of the LSTM, LSTM with bidirectional encoder representations from transformers (BERT), and other deep learning models. Also, collect more data to perform sentiment analysis on the real-world problem.

REFERENCES

[1] K. Sangeetha and D. Prabha, "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*, vol. 12, no. 3, pp. 4117–4126, 2021, doi: 10.1007/s12652-020-01791-9.

[2] J. Zhou and J. Ye, “Sentiment analysis in education research: a review of journal publications,” *Interactive Learning Environments*, vol. 31, no. 3, pp. 1252–1264, Apr. 2023, doi: 10.1080/10494820.2020.1826985.

[3] M. Atif, “An enhanced framework for sentiment analysis of students’ surveys: Arab Open University business program courses case study,” *Business and Economics Journal*, vol. 9, no. 1, 2018, doi: 10.4172/2151-6219.1000337.

[4] C. A. Pacol and T. D. Palaoag, “Enhancing sentiment analysis of textual feedback in the student-faculty evaluation using machine learning techniques,” *European Journal of Engineering Science and Technology*, vol. 4, no. 1, pp. 27–34, 2021, doi: 10.33422/ejest.v4i1.604.

[5] K. Z. Aung and N. N. Myo, “Sentiment analysis of students’ comment using lexicon-based approach,” *Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017*, pp. 149–154, 2017, doi: 10.1109/ICIS.2017.7959985.

[6] H. Newman and D. Joyner, “Sentiment analysis of student evaluations of teaching,” *Artificial Intelligence in Education: 19th International Conference, AIED 2018*, pp. 246–250, 2018, doi: 10.1007/978-3-319-93846-2_45.

[7] Q. Rajput, S. Haider, and S. Ghani, “Lexicon-based sentiment analysis of teachers’ evaluation,” *Applied Computational Intelligence and Soft Computing*, pp. 1–12, 2016, doi: 10.1155/2016/2385429.

[8] S. Rani and P. Kumar, “A sentiment analysis system to improve teaching and learning,” *Computer*, vol. 50, no. 5, pp. 36–43, 2017, doi: 10.1109/MC.2017.133.

[9] F. F. Balahadia, M. C. G. Fernando, and I. C. Juanatas, “Teacher’s performance evaluation tool using opinion mining with sentiment analysis,” in *Proceedings - 2016 IEEE Region 10 Symposium, TENSYMP 2016*, 2016, pp. 95–98, doi: 10.1109/TENCONSpring.2016.7519384.

[10] M. Sivakumar and U. S. Reddy, “Aspect based sentiment analysis of students opinion using machine learning techniques,” in *Proceedings of the International Conference on Inventive Computing and Informatics, ICICI 2017*, 2018, pp. 726–731, doi: 10.1109/ICICI.2017.8365231.

[11] N. Capuano, S. Caballé, J. Conesa, and A. Greco, “Attention-based hierarchical recurrent neural networks for MOOC forum posts analysis,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 11, pp. 9977–9989, 2021, doi: 10.1007/s12652-020-02747-9.

[12] Z. Kastrati, A. S. Imran, and A. Kurti, “Weakly supervised framework for aspect-based sentiment analysis on students’ reviews of MOOCs,” *IEEE Access*, vol. 8, pp. 106799–106810, 2020, doi: 10.1109/ACCESS.2020.3000739.

[13] A. Onan, “Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach,” *Computer Applications in Engineering Education*, vol. 29, no. 3, pp. 572–589, 2021, doi: 10.1002/cae.22253.

[14] M. Elbawab and R. Henriques, “Machine learning applied to student attentiveness detection: using emotional and non-emotional measures,” *Education and Information Technologies*, vol. 28, no. 12, pp. 15717–15737, 2023, doi: 10.1007/s10639-023-11814-5.

[15] A. Baqach and A. Battou, “A new sentiment analysis model to classify students’ reviews on MOOCs,” *Education and Information Technologies*, pp. 1–28, 2024, doi: 10.1007/s10639-024-12526-0.

[16] B. P. Bhagat and S. S. D. -Dandge, “A comparison of different machine learning techniques for sentiment analysis in education domain,” *Cognitive Science and Technology*, pp. 441–450, 2023, doi: 10.1007/978-981-19-8086-2_43.

[17] N. Altrabsheh, M. Cocea, and S. Fallahkhair, “Sentiment analysis: towards a tool for analysing real-time students feedback,” *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, vol. 2014-Decem, pp. 419–423, 2014, doi: 10.1109/ICTAI.2014.70.

[18] G. S. Chauhan, P. Agrawal, and Y. K. Meena, “Aspect-based sentiment analysis of students’ feedback to improve teaching–learning process,” *Smart Innovation, Systems and Technologies*, vol. 107, pp. 259–266, 2019, doi: 10.1007/978-981-13-1747-7_25.

[19] G. G. Esparza *et al.*, “A sentiment analysis model to analyze students’ reviews of teacher performance using support vector machines,” *Advances in Intelligent Systems and Computing*, vol. 620, pp. 157–164, 2018, doi: 10.1007/978-3-319-62410-5_19.

[20] G. Gutiérrez, J. Ponce, A. Ochoa, and M. Álvarez, “Analyzing students reviews of teacher performance using support vector machines by a proposed model,” *Communications in Computer and Information Science*, vol. 820, pp. 113–122, 2018, doi: 10.1007/978-3-319-76261-6_9.

[21] J. A. P. Lalata, B. Gerardo, and R. Medina, “A sentiment analysis model for faculty comment evaluation using ensemble machine learning algorithms,” in *ACM International Conference Proceeding Series*, 2019, pp. 68–73, doi: 10.1145/3341620.3341638.

[22] C. P. -Inwong and W. S. Rungworawut, “Teaching senti-lexicon for automated sentiment polarity definition in teaching evaluation,” in *Proceedings - 2014 10th International Conference on Semantics, Knowledge and Grids, SKG 2014*, 2014, pp. 84–91, doi: 10.1109/SKG.2014.25.

[23] W. Kaur, V. Balakrishnan, and B. Singh, “Improving teaching and learning experience in engineering education using sentiment analysis techniques,” *IOP Conference Series: Materials Science and Engineering*, vol. 834, no. 1, 2020, doi: 10.1088/1757-899X/834/1/012026.

[24] N. K. Singh, D. S. Tomar, and A. K. Sangaiah, “Sentiment analysis: a review and comparative analysis over social media,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 1, pp. 97–117, 2020, doi: 10.1007/s12652-018-0862-8.

[25] P. P. Kumar, P. Gurunadh, and B. Maram, “A study on sentimental analysis using machine learning approach,” *Journal of Advanced Research in Applied Artificial Intelligence and Neural Network*, vol. 5, no. 1, pp. 1–4, 2021.

[26] Y. Chandra and A. Jana, “Sentiment analysis using machine learning and deep learning,” in *Proceedings of the 7th International Conference on Computing for Sustainable Global Development, INDIACom 2020*, 2020, pp. 1–4, doi: 10.23919/INDIACom49435.2020.9083703.

[27] J. G. K. Mabunda, A. Jadhav, and R. Ajoodha, “Sentiment analysis of student textual feedback to improve teaching,” *Interdisciplinary Research in Technology and Management*, pp. 643–651, 2021, doi: 10.1201/9781003202240-100.

[28] Z. Nasim, Q. Rajput, and S. Haider, “Sentiment analysis of student feedback using machine learning and lexicon based approaches,” in *International Conference on Research and Innovation in Information Systems, ICRIIS*, 2017, pp. 1–6, doi: 10.1109/ICRIIS.2017.8002475.

[29] S. Qaiser, N. Yusoff, M. Remli, and H. K. Adli, “A comparison of machine learning techniques for sentiment analysis,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 3, pp. 1738–1744, 2021, doi: 10.17762/turcomat.v12i3.999.

[30] V. D. Nguyen, K. V. Nguyen, and N. L. T. Nguyen, “Variants of long short-term memory for sentiment analysis on Vietnamese students’ feedback corpus,” *Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering, KSE 2018*, pp. 306–311, 2018, doi: 10.1109/KSE.2018.8573351.

[31] R. Z. Cabada, M. L. B. Estrada, and R. O. Bustillos, “Mining of educational opinions with deep learning,” *Journal of Universal Computer Science*, vol. 24, no. 11, pp. 1604–1626, 2018.

[32] I. Sindhu, S. M. Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, “Aspect-based opinion mining on student’s feedback for faculty teaching performance evaluation,” *IEEE Access*, vol. 7, pp. 108729–108741, 2019, doi: 10.1109/ACCESS.2019.2928872.

[33] A. Abdi, G. Sedrakyan, B. Veldkamp, J. V. Hillegersberg, and S. M. V. D. Berg, “Students feedback analysis model using deep learning-based method and linguistic knowledge for intelligent educational systems,” *Soft Computing*, vol. 27, no. 19, pp. 14073–14094, 2023, doi: 10.1007/s00500-023-07926-2.

[34] I. A. Kandhro, S. Wasi, K. Kumar, M. Rind, and M. Ameen, “Sentiment analysis of students’ comment using long-short term model,” *Indian Journal of Science and Technology*, vol. 12, no. 8, pp. 1–16, 2019, doi: 10.17485/ijst/2019/v12i8/141741.

[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv-Computer Science,* pp. 1-12, 2013, doi: 10.48550/arXiv.1301.3781.

[36] L. Breiman, “Random forests,” in *Machine learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[37] K. O’shea and R. Nash, “An introduction to convolutional neural networks,” *arXiv-Computer Science,* pp. 1-11, 2015, doi: 10.48550/arXiv.1511.08458.

[38] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[39] A. Vaswani *et al.*, “Attention is all you need,” *31st Conference on Neural Information Processing Systems (NIPS 2017),* Long Beach, CA, USA, vol. 30, pp. 1-11, 2017.

[40] H. Abbasimehr and R. Paki, “Improving time series forecasting using LSTM and attention models,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 1, pp. 673–691, 2022, doi: 10.1007/s12652-020-02761-x.

[41] B. Liu, “Text sentiment analysis based on CBOW model and deep learning in big data environment,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 2, pp. 451–458, 2020, doi: 10.1007/s12652-018-1095-6.

BIOGRAPHIES OF AUTHORS

Bhavana P. Bhagat working as Sr. Lecturer in Department of Computer Engineering Government Polytechnic Yavatmal (M.S.), India. She is Research Scholar in Department of Computer Science and Engineering Sipna College of Engineering and Technology, Amravati, India and affiliated to Sant Gadge Baba Amravati University Amravati (M.S.), India. She has 26 years of teaching experience. Her area of specialization is artificial intelligence, machine learning, and natural languge processing. She has presented papers in 4 international and 2 national conferences. She can be contacted at email: bhavana.bhagat30475@gmail.com.

Sheetal S. Dhande-Dandge working as Professor at Department of Computer Science Engineering and Technology Amravati. She has Completed Ph.D. from Professor Ram Meghe Institute of Research and Technology Amravati (M.S.), India affiliated to Sant Gadge Baba Amravati University Amravati. (M.S.), India. She has 24 years of teaching and experience. Her area of specialization is artificial intelligence. She has published over 63 papers in international journals and conferences. She can be contacted at email: sheetaldhandedandge@gmail.com.

Sandeep R. Raut working as IB Computer Science teacher in Shanghai Qibao Dwight High School, Shanghai, China. He has 14 years of teaching experience. He participated recently in “Going deeper with technology in education and social media" (EARCOS)”, and “Teaching ESL students in mainstream classrooms (TESMC)”, which focuses on how to engage and support students, by taking into consideration their social, developmental, and language needs. He can be contacted at email: sraut@qibaodwight.org.