IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 2, June 2024, pp. 2173∼2184
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i2.pp2173-2184

Sentiment analysis of student feedback using attention-based RNN and transformer embedding

Imad Zyout[1,2], Mo'ath Zyout[3]
[1]Department of Computer and Communication Engineering, College of Engineering, Tafila Technical University, Tafila, Jordan
[2]Department of Engineering Technology and Science, Faculty of Engineering, Higher Colleges of Technology, Abu Dhabi, UAE
[3]Ministry of Education, Amman, Jordan

Article Info

Article history:

Received Jun 25, 2023
Revised Nov 23, 2023
Accepted Dec 2, 2023

Keywords:

Deep learning
Embedding techniques
Natural language processing
Sentiment analysis
User satisfaction

ABSTRACT

Sentiment analysis systems aim to assess people's opinions across various domains by collecting and categorizing feedback and reviews. In our study, researchers put forward a sentiment analysis system that leverages three distinct embedding techniques: automatic, global vectors (GloVe) for word representation, and bidirectional encoder representations from transformers (BERT). This system features an attention layer, with the best model chosen through rigorous comparisons. In developing the sentiment analysis model, we employed a hybrid dataset comprising students' feedback and comments. This dataset comprises 3,820 comments, including 2,773 from formal evaluations and 1,047 generated by ChatGPT and prompting engineering. Our main motivation for integrating generative AI was to balance both positive and negative comments. We also explored recurrent neural network (RNN), gated recurrent unit (GRU), long short-term memory (LSTM), and bidirectional long short-term memory (Bi-LSTM), with and without pre-trained GloVe embedding. These techniques produced F-scores ranging from 67% to 69%. On the other hand, the sentiment model based on BERT, particularly its KERAS implementation, achieved higher F-scores ranging from 83% to 87%. The Bi-LSTM architecture outperformed other models and the inclusion of an attention layer further enhanced the performance, resulting in F-scores of 89% and 88% from the Bi-LSTM-BERT sentiment models, respectively.

Corresponding Author:

Imad Zyout
Department of Computer and Communication Engineering, College of Engineering
Tafila Technical University
Tafila, Jordan
Email: izyout@ttu.edu.jo

1. INTRODUCTION

In the realm of natural language processing (NLP), sentiment analysis (SA) serves as a crucial tool developed to extract the emotional content hidden within user feedback [1], [2]. SA algorithms have been developed to handle text data collected from various sources, including social media, educational online sites, healthcare records, and student reviews [3]. Feedback represents the response of end-users, whether online or offline, regarding provided services or their level of satisfaction [4]. SA techniques play a vital role in the development and enhancement of both commercial [5] and educational services [6]. User feedback, whether provided online or offline, reflects their satisfaction with the services received, particularly in the realm of education. SA models have been employed to examine course evaluations and student feedback regarding their satisfaction with courses and teachers. Similarly, SA can be employed in other settings where users offer feedback on electronic services [5]. Analyzing reviews by SA is important for service providers to identify areas for improvement and enhance the overall user experience, ultimately resulting in higher user satisfaction [7], [8]. The application of SA in the educational sector is mainly to enhance the learning experience for both teachers and students [9]. With the advent of advanced educational services and distance learning platforms, the integration of SA into the field of education has become increasingly relevant [10]. SA systems are built to collect and analyze students' responses to evaluate their satisfaction with various aspects of the education subject, including the teacher, assignments, and exams [11].

The process of SA and the extraction of feelings, from the text, is done in the form of different levels. These levels include entity level, where sentiment or opinion analysis is conducted on feedback related to an educational entity [12]; sentence level, where the sentiment of the document towards a particular context is analyzed to ascertain if it is positive or negative [13]; document level, where it is determined if the document expresses a positive or negative sentiment towards a given context [14]; and aspect level, providing insight into the positive or negative aspects of educational practices [15]. Particularly, in the application of SA to the education sector, educators can gain valuable insights into student sentiment, which can inform and improve the design and delivery of educational content and services [16].

The key stages to create an effective educational SA model are,as the first stage, the collection and labeling of a feedback dataset [17]. The second stage involves various prepossessing steps, such as cleaning, tokenizing, removing stop words, and stemming. Additionally, the words are encoded using modeling techniques such as term frequency-inverse document frequency (TF-IDF) and term frequency (TF), with the n-gram should be selected before embedding these words into machine learning (ML) algorithms [18], [19]. SA techniques rely on the use of classical supervised ML and modern algorithms to categorize student feedback into positive, negative, or neutral sentiments. Traditional ML algorithms including support vector machines (SVM), decision trees (DT), random forest (RF), and naive Bayes (NB) are often used alone or combined to develop the SA model [3], [19]. In addition, ML techniques such as voting, ensemble, and bagging are employed to improve the accuracy of SA models in the field of education. Furthermore, advanced ML including deep learning techniques like long short-term memory (LSTM) networks [7] and convolutional neural networks (CNN) are utilized to create both supervised and unsupervised SA models for educational purposes [20]. SA approaches that utilize deep learning techniques employ embedding methods to normalize a sequence of vectors into a fixed dimension. The word2vec embedding technique, which utilizes neural networks (NN) to vectorize words into a single vector based on a large corpus of words in a context, has been widely used in SA [21]. Another popular embedding technique is global vectors (GloVe) for word representation, which depends on the co-occurrence of words in a given context [22]. In recent years, embedded models such as bidirectional encoder representations from transformers (BERT)-base [23], robustly optimized bidirectional encoder representations from transformers approach (RoBERTa)-base [24], and a lite bidirectional encoder representations from transformers (ALBERT)-base [25] have been developed based on self-attention mechanisms to retain the position of words in the context. These models obtain vectors by surrounding the words of the objective word and have shown remarkable performance in various NLP tasks, including sentiment analysis [26], [27].

In higher education, the sentiment analysis of student comments, leverages advances in artificial intelligence (AI) and NLP fields to adequately extract constructive feedback and opinions; and learning aspects. A representative SA model can improve and tune teaching methodology to be more effective and to suit different groups of learners. Efforts to build SA models have used different methods to pre-process text and applied various NLP tools including text-cleaning, -tokenizing, and text vectorization methods. Research studies also vary in modeling the opinion-mining problem and in the approach utilized to accomplish the opinion-recognition task. Following the application of different text processing and vectorization methods, the classification problem was solved using divers techniques including classical and modern ML algorithms. Pallathadka *et al.* [28] suggested to forecast the student's performance using ML algorithms including NB, ID3, C4.5, and SVM. On an online dataset, from UCI, to test these four models, the SVM achieved the highest accuracy of 89%. Toçoğlu and Onan *et al.* [29] conducted a sentiment analysis study on Turkish student reviews using various ML algorithms (SVM, NB, logistic regression (LR), RF, AdaBoost, bagging, and the voting algorithm) and text vectorization methods. The study showed that the models based on TF-IDF outperformed the other approaches with scores ranging from 57% to 73%. Okoye *et al.* [30] developed an educational process and data mining plus machine learning (EPDM + ML) model that uses text mining and k-nearest neighbor (KNN) algorithm to analyze teacher performance based on student evaluations. Their analysis showed that 76.4% of the student comments they analyzed were predominantly positive, while 23.6% contained some kind of positive or negative sentiment. The study also found that female students are more likely to recommend teachers based on sentiment, with a precision, recall, specificity, accuracy, and F1-score of 100%, while males are slightly more influenced by emotion with a precision of 94.4%, recall of 100%, specificity of 97.3%, accuracy of 97.3%, and F1-score of 97.1%. The EPDM+ML model was shown to be an effective predictor of student recommendations for teachers with a zero error rate, indicating its potential usefulness in educational settings. Faizi and El Fkihi [31] applied SA to classify positive and negative student reviews collected from course evaluations and social media platforms. Using the SVM algorithm, they achieved an accuracy score of 93.35%. Lalata *et al.* [32] used SA to analyze student comments in the classroom. The ensemble and individual models of LR, SVM, DT, and RF algorithms were evaluated and compared on a dataset of comments of 1413 positive, 327 negative, and 82 neutral comments. The algorithm-based voting method produced the best accuracy of 90.32%. Rakhmanov [33] compared different text vectorization models such as TF-IDF and counter vector for analyzing students' comments. They built a sentiment model and compared several ML algorithms including RF, SVM, NB, gradient boosting, and artificial neural network (ANN). The models were evaluated using 55,000 TF-IDF features extracted from student comments. The results showed that the RF-based TF-IDF method was the most effective, achieving an accuracy of 97%. Sindhu *et al.* [34] utilized multiple feature extraction including TF-IDF, true false (TF), and true positive (TP) with n-grams ranging from 1 to 3 and several ML techniques to analyze massive open online courses (MOOCs) reviews.

Several studies have recently utilized modern ML methods, particularly, deep learning techniques. The first application of deep learning for evaluating faculty teaching performance from students' feedback was presented in [35]. The study presented the supervised aspect-based opinion mining system based on a two-layered LSTM model. On two datasets including a manually tagged dataset and a standard SemEval-2014 dataset and utilizing the domain embedding, the proposed system achieves high accuracy rates of 91% and 93% for both aspect extraction and sentiment polarity detection tasks, respectively. Onan [35] combined CNN, recurrent neural network (RNN), LSTM, and gated recurrent unit (GRU) and three embedding techniques: word2vec, GloVe, and FastText, were using 66,000 MOOC online reviews and the LSTM-based GloVe embedding achieved an accuracy of 95.8%. Yousafzai *et al.* [36] integrated the bidirectional long short-term memory (Bi-LSTM) NN with an attention mechanism and a feature selection method to forecast student performance. The proposed approach attained a 91% accuracy rate on the UCI dataset, utilizing 33 features to characterize a student's behavior throughout a course. A new stemming algorithm was presented in [37] to enhance the SA accuracy of Hausa language, a widely spoken in West Africa. The SA of Hausa language was done classical and modern ML techniques including transformer approaches such as BERT and RoBERTa. Authors developed a monolingual large corpus dataset of about 40,000 Hausa-English comments, named the HESAC. Using the new stemming and achieving about 97.4% by applying the suggested algorithm during the pre-processing phase, which was slightly better than the cross-lingual approach.

This paper presents the SA approach, which is intended to determine the sentiment of student comments about their courses and teachers. The researchers manually collected and labeled comments expressing opinions about courses. To address the limitations of a small dataset of negative comments, the large language model (LLM) and generative AI, specifically ChatGPT, were used to synthesize additional comments. This project utilizes prompt engineering, the design of specific prompts to guide the output of LLM, and generates a wide spectrum of student feedback, including exam question difficulty, teaching style, and fairness of grading. To the best of the author's knowledge, this is the first work, in the context of SA of customer review, that leverages generative AI to build the development dataset. This work also explored several SA model architectures based on deep learning algorithms such as RNN, GRU, LSTM, and Bi-LSTM. To optimize the performance of the proposed SA model, Glove and Bert were utilized to improve the model's performance on the test dataset. The best network was selected through multiple experiments and its performance was optimized using attention layers.

2. METHODS

The methodology for developing the SA approach can be described in three stages. Firstly, the dataset is preprocessed to clean and transform the data. Secondly, the SA models are developed and trained on the preprocessed data. Finally, the performance of the models is evaluated and compared using various metrics. Figure 1 illustrates the three stages of the SA model development process.

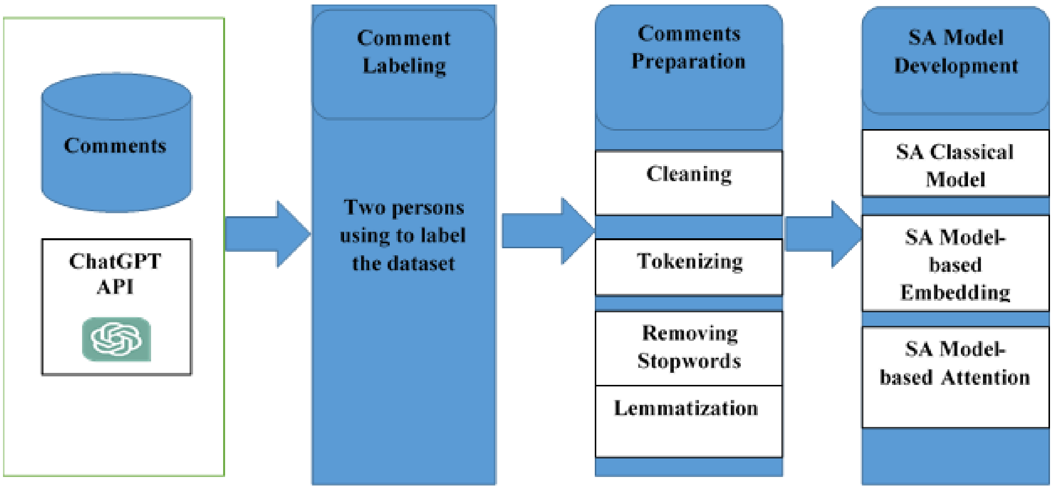

Figure 1. An overview of SA-model architecture

2.1. Data collection and labeling

The initial stage in the development of the SA model encompassed the acquisition of a comprehensive dataset comprising students' comments on diverse courses and instructors. The dataset, consisting of a total of 3820 comments, included 2773 comments provided by human students and 1047 synthetic comments generated by ChatGPT GPT-3.5. Subsequently, two domain experts with more than ten years of experience in the education field manually reviewed and labeled the comments, classifying them into positive and negative sentiments. A comment was categorized as positive if both evaluators confirmed it as such. Conversely, a comment was considered negative if the evaluators marked it as such. To determine whether a student's comment was positive or negative, both keywords and the context of the comment were considered. For example, the following text, ”change your teaching approach, be more creative, and include some fun activities,” which is informal, was labeled by the two experts as negative. Additionally, any comment that received an equal number of positive and negative classifications was eliminated from the dataset due to the insignificance of mixed emotions, accounting for less than 0.001% of the total number of classifications.

2.2. Data preprocessing

The second stage of the proposed SA framework involved preprocessing the comments obtained from the participants, which is shown in Table 1. The preprocessing stage comprised a series of basic steps which included repairing, cleaning, and encoding the comments into vector form. To clean the text in each student's comment, several steps were employed. The efficacy of each step was dependent on specific methods from the natural language toolkit (NTLK) and spacy packages. One of the methods utilized was the named-entity recognition (NER) algorithm from spacy, which was employed to clean the name of the teacher to respect the privacy of each individual. Additionally, special characters such as $*, /, ., ;, :$, and $0 - 9$ were removed from the text, as well as some Arabic words used by the students in their comments. The final step was particularly crucial in addressing frequently occurring and redundant characters in words such as ”bestttttttt,” which was reduced to ”best,” and ”worssssst,” which was transformed into ”worst.” To assess each segment of each statement at the word level, the comments were segmented into a list of words using the NLTK tokenizing algorithm. Additionally, certain words in the comments lacked coherence and were deemed irrelevant to the analysis. To address this, words with a length less than or equal to three were identified and added to a list of stop words such as the, of,. . . , their. These stop words were subsequently removed using the stop words dictionary from the NLTK package. The lemmatization algorithm, in this work, was utilized to transform words into their base form for each comment. This process involved removing the inflectional endings of each word to return it to its base form. By applying lemmatization, the resulting comments were simplified and standardized, facilitating the analysis and interpretation of the data.

Table 1. Examples of raw comments of student feedback on teachers

Comment	Label
He doesn't teach well and brings exams that are very hard not from the PowerPoint	NEG
He doesn't bother himself to explain the information that should be explained by him	NEG
He explains the practical labs very well	POS
He gives most of the students absences, and he gave me an absence when I was in front of him and I wasn't late, but he marked me as late and absent. Why?	NEG
He has an excellent method of teaching	POS
He always helps us	POS

2.3. Sentiment analysis architecture

The final stage of the SA model's design involved utilizing various NLP techniques based on deep learning to select the best SA architecture to achieve superior performance during testing. Initially, the process of extracting feelings from student comments was performed manually by experts and served as a basic plan for preparing the training data used to develop the sentiment model. Several models were built, including popular NN such as LSTM, simple RNN, GRU, and Bi-LSTM. In addition, two text embedding techniques, base-BERT and GloVe, were used to encode words into a robust sentiment model that could extract emotions from student comments. Attention layers were incorporated to improve the performance of the SA model. The combination of these techniques led to the development of a sophisticated SA model capable of accurately detecting emotions in student comments.

2.3.1. Embedding based on global vector

GloVe embedding is a technique used to learn word vectors where the objective of the training is to obtain vectors such that the dot product of any two vectors is equivalent to the logarithm of the probability of the two corresponding words appearing together [22]. This association establishes a connection between the logarithmic ratios of co-occurrence probabilities and vector disparities in the word vector space [6], [19]. By leveraging this relationship, GloVe embedding produces vectors that excel at capturing semantic relationships among words. Figure 2 illustrates the steps involved in generating GloVe word embeddings.

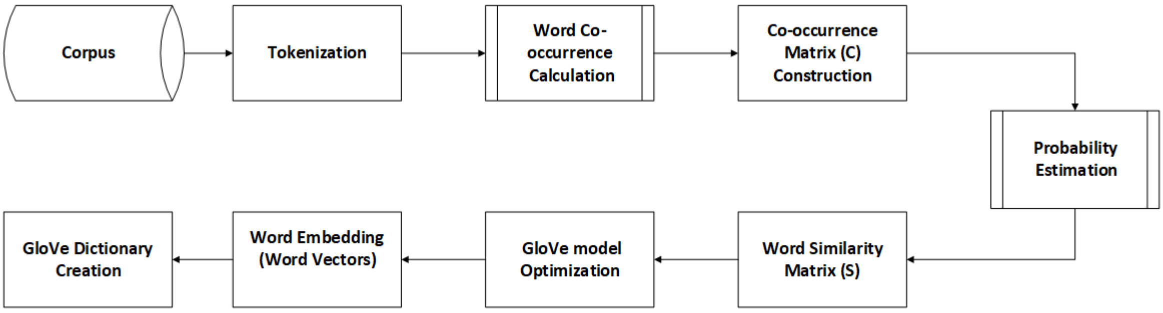
Corpus
Tokenization
Word Co-occurrence Calculation
Co-occurrence Matrix (C) Construction
Probability Estimation
Word Similarity Matrix (S)
GloVe model Optimization
Word Embedding (Word Vectors)
GloVe Dictionary Creation

Figure 2. GloVe embedding steps

2.3.2. Embedding based on Keras BERT

The Keras BERT layer is a pre-trained language model that utilizes the BERT architecture to create word embeddings that capture contextual information. The Keras BERT layer has been pre-trained on a large volume of text data and can encode words into multi-dimensional vectors that reflect the context in which they are used. The resulting embeddings are rich in semantic information and context-dependent. The layer can be further optimized on a smaller, task-specific dataset to enhance its performance on particular tasks [23].

Figure 3 illustrates the process of embedding, self-attention, and output tokens in Keras-BERT. It shows how the input tokens are transformed into embedded representations, followed by the self-attention mechanism that captures the dependencies and relationships between the tokens. The attention weights are calculated based on the query and value vectors, which are then used to obtain a weighted sum of the value vectors [23]. The final output tokens represent the processed and transformed representations of the input tokens, ready for further processing or downstream tasks.

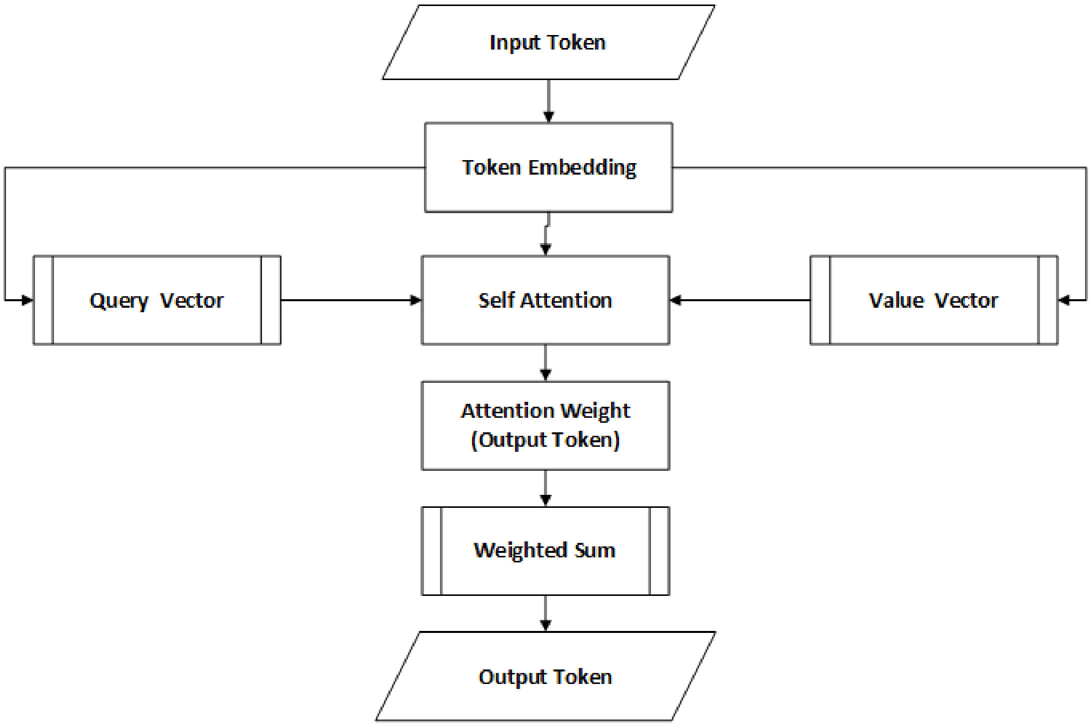

Figure 3. Keras-BERT embedding attention-alt

The architecture of the SA-classical and SA-GloVe models utilizes a variety of multi-NN, including simple RNN, GRU, LSTM, and Bi-LSTM. Two embedding approaches are adopted: one with an embedded vector space having random weights (max length 500), and the other using initial weights from the GloVe dictionaries. Each model consists of a single layer with 100 neurons and integrates both dropout and spatial dropout-1D to minimize overfitting. Post feature extraction in a 100-dimensional space, the softmax activation function determines the probability ratio between positive and negative classes.

The SA-RNN model had an embedding layer with parameters (word length, 300, embedding matrix, 500), SpatialDropout1D of 0.3, RNN layer with 100 neurons, and a Dense layer with 2 neurons. The SA-GRU model had an embedding layer with parameters (word length, 300, embedding matrix, 500), SpatialDropout1D of 0.3, a GRU layer with 100 neurons, and a dense layer with 2 neurons. The SA−LSTM model had an embedding layer with parameters (word length, 300, embedding matrix, 500), Dropout of 0.25, LSTM layer with 100 neurons, and a dense layer with 2 neurons. Finally, the SA-Bi−LSTM model had an embedding layer with parameters (word length, 300, embedding matrix, 500), dropout and recurrent dropout of 0.3 and 0.25, respectively, a Bi−LSTM layer with 100 neurons, and a dense layer with 2 neurons.

Table 2 displays the second set of SA models, utilizing Keras' BERT implementation. These models employ NN (RNN, GRU, LSTM, and Bi-LSTM) similar to classical SA models but differ in using BERT layers for word embeddings. BERT, pre-trained on extensive text data, encodes complex linguistic relationships. Each BERT-based SA model includes one layer with dropout and SpatialDropout1D to prevent overfitting. Extracted features (100 dimensions) from reviews are input to prediction layers using softmax activation. Overall, BERT-based SA models exhibit superior performance compared to classical models, highlighting the advantages of utilizing pre-trained language models like BERT for text classification tasks.

2.3.3. Sentiment analysis-with attention layer parameters

In this section, we describe two methods of SA that incorporate attention layers. The SA-Bi-LSTM and SA-Bi-LSTM-bert models were developed by adding an attention layer before the prediction layers. The parameters of each layer are listed in Table 3. The number of neurons in the embedding layer and the BERT embedding layer are the same, but the attention layer differs from other layers by receiving the output encoded features of the Bi-LSTM layer with a shape of 200 dimensions. The attention layer is responsible for calculating the weights for input features of the previous Bi-LSTM layer. To do this, the features are normalized between

-1 and 1 using the Tanh activation function and then computed using the softmax activation function to create a set of weights. These weights are used to create context vectors by concatenating the input data with the output weights and multiplying them with the input features. The resulting vectors are then added together to create the final form of the context vector. Both SA-Bi-LSTM and SA-Bi-LSTM-BERT models use this attention layer to optimize the prediction scores for the two approaches. The output of the attention layer is input into a dense layer to find the prediction score for each class.

Table 2. Sentiment analysis model architectures with BERT embedding

Model	Layer	Parameters
SA-RNN	Embedding	BERT_layer
	SpatialDropout1D	0.3
	RNN	100
	Dense	2
SA-GRU	Embedding	BERT_layer
	SpatialDropout1D	0.3
	GRU	100
	Dense	2
SA-LSTM	Embedding	BERT_layer
	Dropout	0.25
	LSTM	100
	Dense	2
SA-Bi-LSTM	Embedding	BERT_layer
	Dropout, recurrent_dropout	0.3, 0.25
	Bi-LSTM	100
	Dense	2

Table 3. Sentiment analysis model architectures with attention layer

Model	Layer	Parameters
Bi-LSTM-attention	Embedding	(word length, 300, EM, 500)
	Bidirectional	100
	Dropout rate	0.3
	Attention	Output 200
	Dense	2
Bi-LSTM-bert-attention	Embedding	BERT layer
	Bi-LSTM	100
	Dropout rate	0.3, 0.3
	Attention	Output 200
	Dense	2

3. RESULTS AND DISCUSSION

Three main experiments were conducted to assess proposed approaches for enhancing SA of student feedback. The first experiment compared RNN, GRU, LSTM, and Bi-LSTM networks with and without GloVe embeddings, ultimately identifying Bi-LSTM as the best performer. The second experiment utilized the Keras BERT layer to enhance the performance of the Bi-LSTM network. The third experiment introduced attention layers to the best-performing BiLSTM networks with Glove embeddings and the Keras BERT layer, resulting in further improvements in the SA model.

For model development and evaluation, the dataset was split into development and testing datasets. The development dataset, representing 80% of the data, was used for training and validation, while the remaining 20% was allocated for testing. The hyperparameter settings for training various models included a batch size of 64, 20 epochs, a learning rate of 0.001, and the use of the Adam algorithm for parameter optimization. These settings were determined after conducting multiple random experiments to optimize the SA-model learning during training. Additionally, the early stopping technique was employed to address model overfitting and optimize performance metrics such as precision, recall, and F1-score. In the subsequent section, we present the results of our experiments with various SA models, including classical NN and those incorporating pretrained embeddings like GloVe and BERT. Each model underwent training and testing on our student chat dataset, with performance evaluation based on precision, recall, and F1-score metrics.

3.1. Sentiment analysis-classical network

Results of evaluating the four different SA models (SA-RNN, SA-GRU, SA-LSTM, and SA-Bi-LSTM) are presented in Table 4. Overall, SA-Bi-LSTM outperformed all by achieving the highest precision, recall, F-score, and accuracy scores. Specifically, the SA-Bi-LSTM model achieved a precision score of 68%, meaning that it was correct in identifying positive sentiment instances 68% of the time. Specifically, the SA-Bi-LSTM model achieved a precision score of 68%, meaning that it was correct in identifying positive sentiment instances 68% of the time. Additionally, the F-score of SA-Bi-LSTM was 66%, indicating that it was effective in identifying positive sentiment instances while minimizing false positives. The accuracy score of SA-Bi-LSTM was also the highest among the models at 68%. It is important to note that the performance differences between the models were relatively small, and other factors, such as the size and composition of the dataset or the specific parameters used in each model, may have influenced the results. Future research could be conducted to explore the impact of these factors on sentiment analysis model performance.

Table 4. Performance of SA models based-classical network

SA Model	Precision (%)	Recall (%)	F-score (%)	Accuracy (%)
SA-RNN	64	67	65	67
SA-GRU	62	66	64	66
SA-LSTM	63	64	64	64
SA-Bi-LSTM	68	66	66	68

3.2. Sentiment analysis-global vector embedding

Table 5 shows the performance of four sentiment analysis models - SA-RNN-GloVe, SA-GRU-GloVe, SA-LSTM-GloVe, and SA-Bi-LSTM-GloVe - that were trained using GloVe embeddings on the same dataset of customer reviews. The models were evaluated using precision, recall, accuracy, and F-score metrics. The SA-Bi-LSTM-GloVe model achieved the highest scores across all four metrics with precision of 69%, recall of 69%, accuracy of 69%, and F-score of 69%. The SA-RNN-GloVe model had the second-highest scores with precision of 66%, recall of 69%, accuracy of 67%, and F-score of 69%.

Table 5. Performance of SA models with pre-trained Glov embedding

SA Model	Precision (%)	Recall (%)	Accuracy (%)	F-score (%)
SA-RNN-Glov	66	69	67	69
SA-GRU-Glov	64	68	65	68
SA-LSTM-Glov	65	69	67	69
SA-BiL-STM-Glov	69	69	69	69

Overall, the results suggest that using GloVe embeddings can improve the performance of SA models. However, it's important to note that the differences in performance between the models were relatively small and other factors such as the size and composition of the dataset, or the specific parameters used in each model, could also have an impact on performance.

3.3. Sentiment analysis-BERT embedding method

Table 6 shows the performance produced by the four SA models with BERT embeddings: SA-RNN-BERT, SA-GRU-BERT, SA-LSTM-BERT, and SA-Bi-LSTM-BERT. Overall, the models performed very well, with all models achieving high precision, recall, accuracy, and F-score scores. The SA-Bi-LSTM-BERT model had the highest precision, recall, accuracy, and F-score scores, all of which were 86.4%, 86.3%, 86.4%, and 86.5% respectively. The SA-LSTM-BERT and SA-GRU-BERT models also had high scores across all performance metrics, with the SA-RNN-BERT model achieving slightly lower scores in precision, recall, and F-score.

Table 6. Performance of SA models with pre-trained BERT embedding

SA Model	Precision (%)	Recall (%)	Accuracy (%)	F-score (%)
SA-RNN-BERT	85	83	84	83
SA-GRU-BERT	86	86	86	86
SA-LSTM-BERT	85.8	85.8	86	86
SA-Bi-LSTM-BERT	86.5	86.3	86.4	86.5

The high performance of these models is primarily attributed to the use of BERT embedding. BERT models were pre-trained on a large corpus of text data and have shown to be highly effective in NLP tasks. Our results suggest that the SA model, utilizing BERT embedding, excels in accurately identifying sentiment in students' reviews. This achievement has important implications for educational institutes, particularly regarding course reviews and instructors' training.

3.4. Sentiment analysis-with attention layer

Table 7 displays the performance results of two sentiment analysis models employing attention mechanisms: SA-attention-Bi-LSTM and SA-attention-Bi-LSTM-BERT. The results demonstrate that integrating attention mechanisms into SA models has led to enhanced performance. Notably, the SA-attention-Bi-LSTM-BERT model, which incorporates both attention and BERT embeddings, achieved the highest scores among the models evaluated in this study. Nevertheless, the performance differences between the two attention-based models were relatively small. It is essential to acknowledge that factors such as dataset size, composition, or model parameters could also influence the model's performance.

Table 7. Performance evaluation of SA models- BiLSTM with attention

SA model	Precision (%)	Recall (%)	Accuracy (%)	F-score (%)
SA-attention-Bi-LSTM	88	87	88	87
SA-attention-Bi-LSTM-BERT	89	88	89	89

3.5. Sentiment analysis model comparative analysis

This study evaluated three sentiment analysis models Figure 4 on the dataset of student reviews: SA-GloVe embedded models, SA-BERT embedded models, and SA-attention models. SA-Bi-LSTM-GloVe, from the SA-GloVe category, achieved the highest performance with 69% accuracy and F-score, while the other three models scored between 65% and 68%. SA-BERT embedded models, with four architectures, achieved F-scores ranging from 83% to 87%, with SA-Bi-LSTM-BERT attaining the highest F-score of 87%. SA-attention models, including SA-attention-Bi-LSTM and SA-attention-Bi-LSTM-BERT, scored F-scores of 86% and 89%, surpassing SA-GloVe models. Notably, SA-attention-Bi-LSTM-BERT outperformed SA-attention-Bi-LSTM. Comparing results, attention mechanisms, and BERT embeddings improved SA model performance. SA-Bi-LSTM-GloVe excelled among SA-GloVe models, SA-Bi-LSTM-BERT led SA-BERT models, and SA-attention-Bi-LSTM-BERT achieved the highest F-score overall.

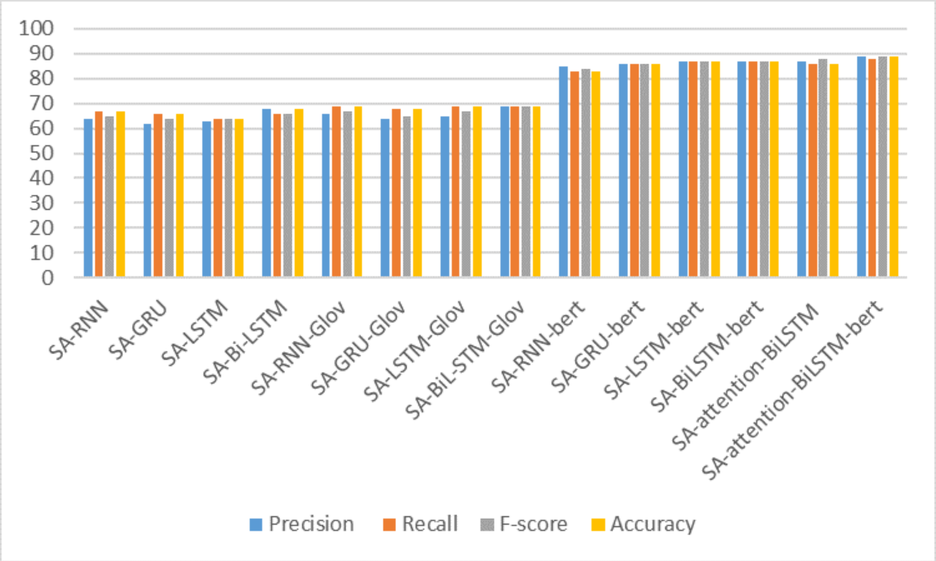

Figure 4. Comparison of all suggested SA models architecture

The SA model, lacking an embedding method, exhibited significant inaccuracies, misclassifying instances like "maybe need to chill a bit" as positive, contrary to annotators and BERT-based SA models. The phrase ”found it hard to complete the assignment on time” was inaccurately labeled by classical networks and GloVe embedding but correctly identified as negative by BERT-based SA. Optimization attempts for GloVe within SA models, including LSTM, GRU, Bi-LSTM, and RNN, aimed to enhance performance but fell short of overcoming inherent limitations, resulting in non-negligible error rates. Challenges in accurately classifying phrases, like ”the teacher seems easily frustrated,” persisted. Notable improvement was seen in the attention layer of BERT-based SA, as illustrated in Figure 5. SA models (Bi-LSTM-attention) and (Bi-LSTM-BERT-attention) outperformed, achieving exceptional predictions with 155 and 160 instances correctly classified in the negative class, and 525 and 532 instances in the positive class, respectively.

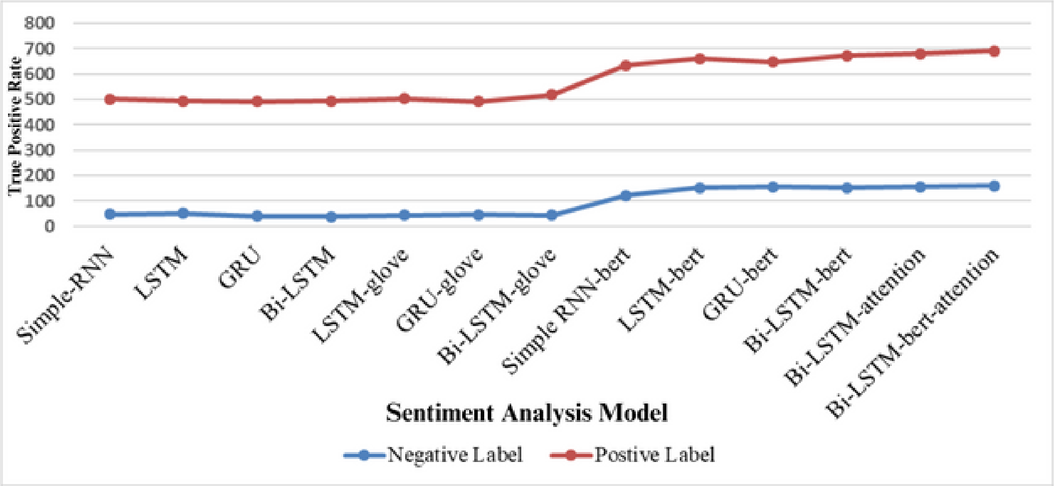

Figure 5. Total of TP of all suggested SA models

4. CONCLUSION

The use of SA has become increasingly popular in analyzing people’s opinions across various domain mains. In the educational context, SA can be used to gauge student satisfaction and demand for teachers’ services. However, such models face limitations due to small database sizes and the diverse deep learning tools used. To overcome this, this proposed a SA system using three methods, including automatic embedding and GloVe and BERT embedding, to analyze a database of student chat categorized into positive and negative feedback. Classical networks such as RNN, GRU, LSTM, and Bi-LSTM were employed, with and without pre-trained GloVe embeddings, achieving F-scores ranging from 67% to 69%. A BERT Keras embedding layer-based sentiment model was also evaluated, achieving F-scores ranging from 83% to 87%. The addition of an attention layer to the Bi-LSTM SA model yielded the highest performance, resulting in an enhanced F-score of 89% for the Bi-LSTM-BERT sentiment model and 88% for the Bi-LSTM sentiment model. The proposed system can thus provide valuable insights into student satisfaction and demand for teachers’ services, contributing to the enhancement of educational quality. Though the results of this work revealed that the SA model architecture, indeed, has influenced the overall performance, the data-centric approach necessitates further research into its potential to boost the model performance.

REFERENCES

[1] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, “Sentiment analysis and opinion mining on educational data: A survey,” *Natural Language Processing Journal*, vol. 2, pp. 1–11, Mar. 2023, doi: 10.1016/j.nlp.2022.100003.

[2] M. Alzyout, E. AL Bashabsheh, H. Najadat, and A. Alaiad, “Sentiment Analysis of Arabic Tweets about Violence Against Women using Machine Learning,” in *2021 12th International Conference on Information and Communication Systems (ICICS)*, 2021, pp. 171–176, doi: 10.1109/ICICS52457.2021.9464600.

[3] D. D. Dsouza, Deepika, D. P. Nayak, E. J. Machado, and N. D. Adesh, “Sentimental analysis of student feedback using machine learning techniques,” *International Journal of Recent Technology and Engineering*, vol. 8, no. 1 Special Issue 4, pp. 986–991, 2019.

[4] S. Ulfa, R. Bringula, C. Kurniawan, and M. Fadhli, “Student Feedback on Online Learning by Using Sentiment Analysis: A Literature Review,” in *2020 6th International Conference on Education and Technology (ICET)*, 2020, pp. 53–58, doi: 10.1109/ICET51153.2020.9276578.

[5] R. Baragash and H. Aldowah, "Sentiment analysis in higher education: A systematic mapping review," *Journal of Physics: Conference Series*, vol. 1860, no. 1, pp. 1–13, 2021, doi: 10.1088/1742-6596/1860/1/012002.

[6] D. K. Dake and E. Gyimah, "Using sentiment analysis to evaluate qualitative students' responses," *Education and Information Technologies*, vol. 28, no. 4, pp. 4629–4647, 2023, doi: 10.1007/s10639-022-11349-1.

[7] I. A. Kandhro, S. Wasi, K. Kumar, M. Rind, and M. Ameen, "Sentiment Analysis of Students Comment by 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.

[8] L. K. Singh, "An Review of Student Sentimental Analysis For Educational Database Using Unsupervised Machine Learning Approaches.," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 9, pp. 2151–2165, 2021.

[9] A. Tzacheva and A. Easwaran, "Emotion Detection and Opinion Mining from Student Comments for Teaching Innovation Assessment," *International Journal of Education (IJE)*, vol. 9, no. 2, pp. 21–32, 2021, doi: 10.5121/ije2021.9203.

[10] 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, 2023, doi: 10.1080/10494820.2020.1826985.

[11] F. Dalipi, K. Zdravkova, and F. Ahlgren, "Sentiment Analysis of Students' Feedback in MOOCs: A Systematic Literature Review," *Frontiers in Artificial Intelligence*, vol. 4, pp. 1–13, 2021, doi: 10.3389/frai.2021.728708.

[12] R. Yang, "Machine Learning and Deep Learning for Sentiment Analysis over Students' Reviews: An Overview Study," *Preprints*, pp. 1-9, 2021.

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

[14] F. S. Dolianiti *et al.*, "Sentiment analysis on educational datasets: a comparative evaluation of commercial tools," *Educational Journal of the University of Patras UNESCO Chair*, vol. 6, no. 1, pp. 262–273, 2019.

[15] T. Shaik *et al.*, "A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis," *IEEE Access*, vol. 10, pp. 56720–56739, 2022, doi: 10.1109/ACCESS.2022.3177752.

[16] D. Wehbe, A. Alhammadi, H. Almaskari, K. Alsereidi, and H. Ismail, "UAE e-Learning Sentiment Analysis Framework," in *The 7th Annual International Conference on Arab Women in Computing in Conjunction with the 2nd Forum of Women in Research*, 2021, pp. 1–4, doi: 10.1145/3485557.3485570.

[17] K. Nilanga, M. Herath, H. Maduwantha, and S. Ranathunga, "Dataset and Baseline for Automatic Student Feedback Analysis," in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 2022, pp. 2042–2049.

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

[19] 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.

[20] I. A. Kandhro, M. A. Chhajro, K. Kumar, H. N. Lashari, and U. Khan, "Student Feedback Sentiment Analysis Model Using Various Machine Learning Schemes A Review," *Indian Journal of Science and Technology*, vol. 14, no. 12, pp. 1–9, 2019, doi: 10.17485/ijst/2019/v12i14/143243.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013, pp. 1–12.

[22] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, vol. 1, pp. 4171–4186.

[24] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," in *ICLR 2020 Conference*, 2019, pp. 1–15.

[25] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020, pp. 1–17.

[26] R. Bensoltane and T. Zaki, "Towards Arabic aspect-based sentiment analysis: a transfer learning-based approach," *Social Network Analysis and Mining*, vol. 12, no. 1, 2022, doi: 10.1007/s13278-021-00794-4.

[27] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1743–1753, 2023, doi: 10.1109/TAFFC.2022.3204972.

[28] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, "Classification and prediction of student performance data using various machine learning algorithms," *Materials Today: Proceedings*, vol. 80, pp. 3782–3785, 2023, doi: 10.1016/j.matpr.2021.07.382.

[29] M. A. Toçoğlu and A. Onan, "Sentiment Analysis on Students' Evaluation of Higher Educational Institutions," in *Intelligent and Fuzzy Techniques: Smart and Innovative Solutions*, Cham: Springer, 2021, pp. 1693–1700, doi: 10.1007/978-3-030-51156-2_197.

[30] K. Okoye, A. Arrona-Palacios, C. Camacho-Zuñiga, J. A. G. Achem, J. Escamilla, and S. Hosseini, "Towards teaching analytics: a contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification," *Education and Information Technologies*, vol. 27, no. 3, pp. 3891–3933, 2022, doi: 10.1007/s10639-021-10751-5.

[31] R. Faizi and S. El Fkihi, "A Sentiment Analysis Based Approach for Exploring Student Feedback," in *Innovative Technologies and Learning*, Cham: Springer, 2022, pp. 52–59, doi: 10.1007/978-3-031-15273-3_6.

[32] 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.

[33] O. Rakhmanov, "A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments," *Procedia Computer Science*, vol. 178, pp. 194–204, 2020, doi: 10.1016/j.procs.2020.11.021.

[34] I. Sindhu, S. Muhammad 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.

[35] 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.

[36] B. K. Yousafzai *et al.*, “Student-performulator: Student academic performance using hybrid deep neural network,” *Sustainability*, vol. 13, no. 17, pp. 1–21, 2021, doi: 10.3390/su13179775.

[37] O. Rakhmanov and T. Schlippe, “Sentiment Analysis for Hausa: Classifying Students’ Comments,” in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022*, 2022, pp. 98–105.

BIOGRAPHIES OF AUTHORS

Imad Zyout serves as division chair of engineering at Higher Colleges of Technology, UAE, and is an associate professor in Computer and Communication Engineering at Tafila Technical University, Jordan. With a Ph.D. from Western Michigan University, he gained international recognition, including an Australian Endeavour Post-Doctorate Fellowship in 2013. He expertise spans computer vision, featuring feature engineering, and machine learning algorithms. His research is notably dedicated to developing efficient algorithms for mammography image analysis. He can be contacted at email: izyout@ttu.edu.jo.

Mo’ath Zyout is currently an instructor in Computer Science at the Qatar Ministry of Education. He is a recipient of the Yarmouk University bachelor’s degree, obtained in 2010. He received his Master’s degree in Data Science from the Jordan University of Science & Technology, Jordan, in 2022. His research in NLP focusing on sentiment analysis and malware classification. He can be contacted at email: moathzyout@gmail.com.