Enhanced Bidirectional LSTM for Sentiment Analysis of Learners' Posts in MOOCs

Chakir Fri[1]*, Rachid Elouahbi[2], Youssef Taki[3], Ahmed Remaida[4]
Laboratory of Computer Science and Applications-Faculty of Sciences, Moulay Ismail University, Meknes, Morocco[1, 2]
ENSAM Meknes, Moulay Ismail University, Meknes, Morocco[3]
Laboratory of Engineering Sciences-National School of Applied Sciences, Ibn Tofaïl University, Kenitra, Morocco[4]

***Abstract*—Massive Open Online Courses (MOOCs) have transformed digital learning, leading to vast amounts of learner-generated content that reflect user experience and engagement. Accurately classifying sentiment from this content is essential for improving course quality, but remains challenging due to subtle linguistic variation and contextual ambiguity. This study proposes a sentiment analysis approach based on an enhanced Bidirectional Long Short-Term Memory (LSTM) model. The enhancements include the integration of data augmentation and regularization techniques to address overfitting and improve generalization. The model was trained and evaluated on a dataset of 29,604 learner discussion posts from Stanford University MOOCs. Experimental results show that the proposed model achieves an accuracy of 88.54% in classifying sentiments into positive, negative, and neutral classes. These results suggest that the enhanced LSTM model offers a reliable solution for large-scale sentiment classification in online education, with potential applications in learner support, curriculum design, and personalized feedback.**

***Keywords*—MOOCs; Sentiment analysis; deep learning; Bidirectional LSTM; data augmentation; regularization techniques**

I. Introduction

Massive Open Online Courses (MOOCs) have revolutionized digital education by enabling global access to quality learning. With thousands of learners actively participating in MOOC platforms, understanding learner sentiment is essential to improving course quality, fostering engagement, and guiding instructional strategies. Sentiment analysis (SA), a subfield of natural language processing (NLP), has become a powerful tool to extract emotional insights from learners' discussions and feedback.

Numerous studies have applied sentiment analysis in educational settings using machine learning (ML) and deep learning (DL) models. Kastrati et al. [1] used a BiGRU model with Word2Vec embeddings to classify MOOC feedback, enhancing the overall sentiment classification pipeline. Zhang and Zhu [2] fine-tuned BERT on educational data to capture contextual sentiment, showing improved performance on short learner posts. Phan et al. [3] integrated an attention-based deep learning architecture for aspect-based sentiment extraction, aiding in pedagogical refinements. Ortigosa et al. [4] applied lexicon-based sentiment analysis to Facebook data for personalized e-learning, improving adaptive content delivery. Onan [5] demonstrated that CNN models with Word2Vec embeddings outperform traditional machine learning techniques in MOOC sentiment classification. Ramesh et al. [6] modeled emotional cues from MOOC discussions using LSTM with attention mechanisms to predict dropout risks, supporting early intervention strategies. Rani and Kumar [7] developed a sentiment-aware feedback system using rule-based NLP techniques to enhance teaching quality. Alatrash et al. [8] proposed a sentiment-driven recommender system for MOOCs that dynamically adjusts learning materials based on learners' emotions. Zhang and Zhu [9] combined sentiment and content analysis using a hybrid deep learning approach to generate fine-grained profiles of learners in LMOOC platforms.

In addition to the works reviewed above, other efforts also demonstrate the need for more robust and domain-specific sentiment models. Chen et al. [10] introduced a semi-supervised learning model tailored to MOOC forums. Sailunaz and Alhajj [11] studied emotion-aware sentiment modeling using social media data. Kumar et al. [12] proposed a multi-task neural architecture combining sentiment and emotion classification. Zhang et al. [13] integrated attention mechanisms to enhance performance on short text sentiment tasks. Priyadharshini et al. [14] designed a CNN–BiLSTM model that showed strong results on diverse emotional datasets.

While valuable, most prior studies do not incorporate advanced data enhancement techniques such as text augmentation or regularization. These methods can significantly reduce overfitting and compensate for limited annotated data both critical challenges in educational datasets. Additionally, many existing models rely on small or generic datasets, which lack the scale and linguistic diversity of MOOC forums.

To overcome these limitations, we propose a novel scalable sentiment analysis framework leveraging Bidirectional Long Short-Term Memory (BiLSTM) networks, enriched with advanced data augmentation and regularization techniques. Our model is rigorously validated using a large-scale real-world dataset of over 29,000 learner discussion posts from Stanford University MOOCs, classifying sentiments into positive, negative, and neutral categories.

The primary contributions of this study are:

- Development of a BiLSTM-based sentiment analysis framework customized for large-scale MOOC discussions, enhanced with data augmentation and regularization.
- Comprehensive benchmarking against established models to validate the framework's effectiveness.

*Corresponding Author.

- Real-world application to a large educational dataset, confirming scalability and practical relevance.

The remainder of this paper is structured as follows: Section II presents the theoretical basis, outlining key concepts in sentiment analysis and deep learning. Section III describes our methodology, detailing preprocessing, model design, and training procedures. Sections IV and V provides results and discusses their implications. Finally, Section VI concludes the paper and outlines future research directions.

II. THEORETICAL BASIS

Understanding sentiment in educational discussions is critical for evaluating learner satisfaction, identifying disengagement, and adapting course content. This section presents the theoretical background that underpins our work, including core distinctions between emotion and sentiment, followed by the rationale for using deep learning, particularly Bidirectional Long Short-Term Memory (BiLSTM) networks in processing MOOC discussions.

A. Emotion and Sentiment

Emotion is a complex human experience defined as a powerful feeling arising from circumstances, mood, or interpersonal connections [15], often manifesting as brief, intense reactions to specific events [16]. Theories of emotion are divided into neurological, physiological, and cognitive categories [17], including the Evolutionary Theory of Emotion [18], James-Lange theory [19], and Schachter-Singer Theory [20]. Emotions can be gauged through dimensional approaches, like Russell's circumplex model [21], or categorical approaches, such as the six basic emotions [22]. In contrast, sentiment refers to the enduring positive or negative feelings shaping opinions [23], involving a mix of emotions, cognition, and behavior [24]. While emotion and sentiment are distinct, many sentiment analysis systems rely on emotion analysis [25] [26].

B. Deep Learning Models for Sentiment Analysis

A Recurrent Neural Network (RNN) is a class of artificial neural networks specifically designed to process and analyze sequential data. It consists of repeating modules that allow information to persist across time steps. Long Short-Term Memory (LSTM), a specialized type of RNN, was introduced to address the instability issues encountered in traditional RNNs, particularly the vanishing gradient problem, which previously hindered their practical applicability. LSTM networks are capable of learning and exploiting long-term temporal dependencies in sequential data by leveraging internal memory cells. These cells enable the model to retain relevant past information and make predictions based on the contextual dependencies present in the input sequence. A defining feature of LSTM architecture, as opposed to other deep learning models such as Convolutional Neural Networks (CNNs), is the presence of three gating mechanisms: the input gate, forget gate, and output gate. These gates regulate the flow of information by selectively incorporating new input (input gate), discarding irrelevant information (forget gate), and transmitting pertinent data to subsequent time steps (output gate) [27]. A schematic representation of these recurrently connected cells is illustrated in Fig. 1.

The input gate is denoted by i, the output gate by o, and the forget gate by f. The cell state is represented as C, the cell output as h, and the input at a given time step as x. As illustrated in Fig. 2, the structure of the LSTM cell enables it to regulate information flow using these components. The following equations formally define the operations performed within an LSTM cell during each time step:

$$f_t = \sigma\left(W_f \cdot \left[h_{\{t-1\}};\, x_t\right] + b_f\right) \quad (1)$$

$$i_t = \sigma\left(W_i \cdot \left[h_{\{t-1\}};\, x_t\right] + b_i\right) \quad (2)$$

$$\widetilde{C_t} = \tanh\left(W_C \cdot \left[h_{\{t-1\}}; x_t\right] + b_C\right) \quad (3)$$

$$C_t = f_t \cdot C_{\{t-1\}} + i_t \cdot \widetilde{C_t} \quad (4)$$

$$o_t = \sigma\left(W_o \cdot \left[h_{\{t-1\}};\, x_t\right] + b_o\right) \quad (5)$$

$$h_t = o_t \cdot \tanh\left(C_t\right) \quad (6)$$

The matrices W represent the learnable weights associated with each gate, while C denotes the updated cell state. These states are propagated forward through the network, as illustrated in Fig. 2, and the weights are optimized using backpropagation through time. The forget gate plays a crucial role in mitigating overfitting by selectively discarding irrelevant information from previous time steps. This gated architecture and its mechanism for controlling information flow are instrumental in addressing the vanishing gradient problem inherent in traditional RNNs. As a result, LSTM networks are particularly effective for modeling complex, non-stationary sequences.

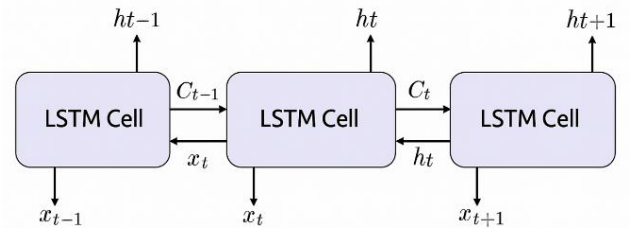

Fig. 1. LSTM Layout with cell connections.

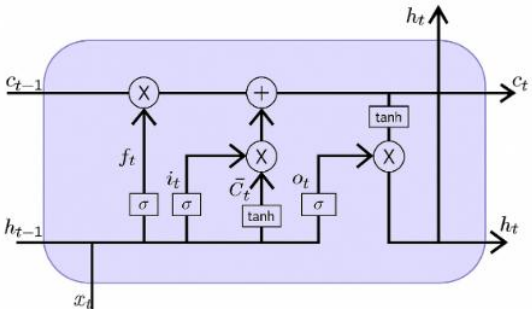

Fig. 2. Architecture of a LSTM cell with various gates.

The standard LSTM model was initially proposed by Hochreiter and Schmidhuber in 1997 [28], and the Bidirectional LSTM (BiLSTM) variant was later introduced by Graves et al. in 2005 [29]. Fig. 3 illustrates the general schematics of LSTM and BiLSTM networks. In an LSTM, each hidden cell receives input influenced by computations performed in cells from preceding time steps. This explicit management of sequential memory makes LSTM particularly suitable for modeling

sequential data. In contrast, the BiLSTM architecture features a bidirectional flow of information, employing two LSTM networks: one processing data in a forward direction, and the other in reverse, with outputs from both networks merging at the output layer. This bidirectional context has been shown to significantly improve accuracy in language modeling [30] [31], and related tasks.

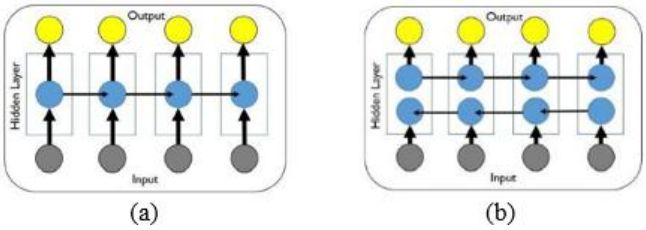

Fig. 3. LSTM (a) and BILSTM (b) Architecture.

In our research, we chose the Bidirectional Long Short-Term Memory (BiLSTM) model due to its effectiveness in capturing contextual dependencies in both forward and backward directions an essential feature for understanding nuanced sentiment in learner-generated content. This makes it particularly suitable for processing the informal, sequential, and often ambiguous language found in MOOC forum posts. Furthermore, many existing models lack mechanisms for addressing challenges such as overfitting, class imbalance, and limited linguistic variability. Our enhanced BiLSTM framework integrates data augmentation and regularization techniques to overcome these limitations and improve generalization across diverse sentiment categories.

III. METHOD

This section presents the methodology adopted for sentiment analysis of learners' posts in the MOOC context. We begin by describing the dataset used in this study, followed by an exploratory analysis to uncover linguistic and sentiment patterns. Subsequently, we discuss the text representation process combining tokenization and pre-trained word embeddings. We then detail the design and training of the Bidirectional Long Short-Term Memory (BiLSTM) model. Finally, we describe the experimental setup, including training parameters and evaluation metrics. The proposed pipeline is illustrated in Fig. 4.

A. Dataset

In this study, we utilized the Stanford MOOC Posts dataset [32], which comprises 29,604 learner forum posts collected from Stanford University's OpenEdX platform between August 2013 and September 2014. The dataset covers six different MOOCs across three academic domains: Education, Medicine, and Statistics. Each post was manually annotated by human coders across several dimensions, including confusion, urgency, opinion, question, answer, and sentiment. Table I summarizes the key metadata of the Stanford MOOC Posts dataset.

The dataset exhibits challenges typical of real-world online text, including class imbalance among sentiment labels, informal expressions, typos, and the use of abbreviations. Recognizing these challenges is critical for effective preprocessing and model design.

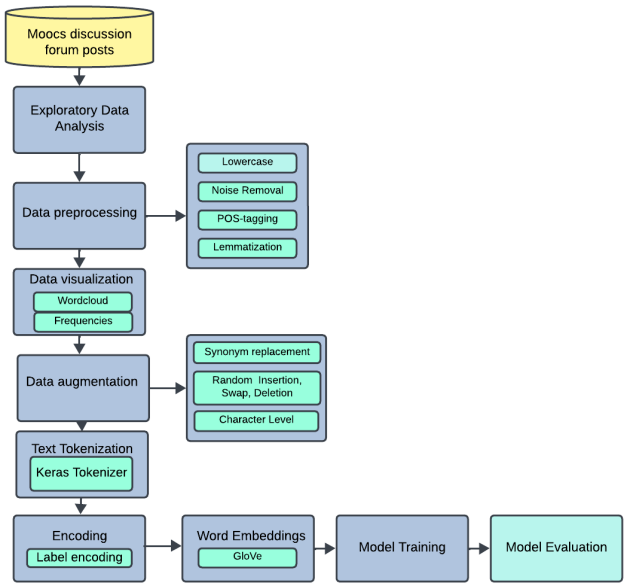

Fig. 4. Proposed methodology.

TABLE I. METADATA OF THE STANFORD MOOC POSTS DATASET

Attribute	Description
Source	Stanford University's OpenEdX platform
Collection Period	August 2013 – September 2014
Number of Courses	6
Number of Posts	29,604
Language	English
Sentiment Labels	1 (Very Negative) to 7 (Very Positive)
Data Fields	Text, Sentiment, Confusion, Urgency, Course Type, Timestamp, Forum Post ID, Forum UID, Anonymized User Info
Post Types	Comment, Comment Thread
Challenges	Class imbalance, informal text, typos, abbreviations

Regarding the sentiment dimension, each post was rated on a 7-point scale, where a score of 7 indicates a highly positive sentiment requiring no instructor intervention, and a score of 1 signifies a highly negative sentiment necessitating immediate instructor attention. This fine-grained labeling provides a valuable resource for sentiment classification tasks.

Table II presents examples of learner posts along with their corresponding sentiment scores.

TABLE II. DATASET

Posts	Score
I am really glad that I entered this MOOC. A lot of interesting things are explained in an engaging manner! Loss of motor control in the cold, the after drop - fantastic!	7
Yes, the parent and teacher do have an important role as an encouraging mentor who continues to learn when to step in and when to step back.	4
TERRIBLE interface design! Just put an obvious 'next' button at the bottom of the main body area or clone the whole linear navigation from the top.	1

The goal of this study was to assess whether a post was positive, negative, or neutral. We considered posts scoring above 4 to be positive, those scoring below 4 to be negative, and those scoring exactly 4 to be neutral.

B. Exploratory Data Analysis

To begin our analysis, we conducted an exploratory study of the dataset. As shown in Fig. 5, the sentiment scores are not evenly distributed across the posts, with a noticeable concentration around the score of 4. Posts labeled with a sentiment score of 4 often exhibit a mixture of positive and negative expressions, making them less straightforward for classification purposes. However, instead of excluding these instances, we retained all posts, including those with a score of 4, to preserve the integrity and representativeness of the dataset.

This decision ensures that our model is exposed to a more realistic distribution of sentiments encountered in real-world learner discussions.

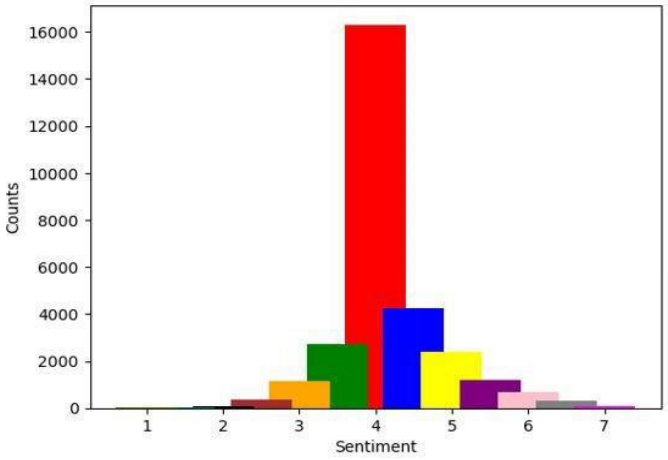

Fig. 5. Sentiment distribution.

Following the initial exploration, we categorized the sentiment scores into three distinct classes to simplify the classification task. Posts with a sentiment score greater than 4 were labeled as positive, those with a score less than 4 as negative, and posts with a score exactly equal to 4 as neutral. The final sentiment distribution after this categorization is illustrated in Fig. 6.

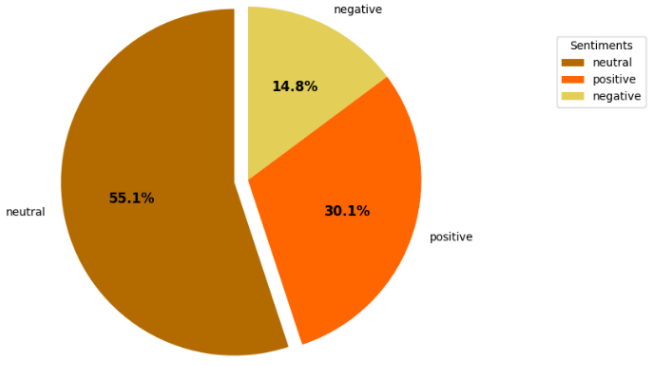

Fig. 6. Final sentiment distribution.

C. Data Preprocessing

In this step, we prepared the textual data by applying a series of preprocessing operations to improve the quality and consistency of the corpus before feeding it into the model. Effective data cleansing is critical in text analysis, as it removes noise and ensures that the inputs are more comprehensible for subsequent natural language processing (NLP) tasks. The following preprocessing procedures were implemented:

1) Data inspection: The data was inspected to identify any missing values or unhelpful data. Any null values and irrelevant columns were dropped. Since the "Post" column is our target data, we retained only the "Post" in the final DataFrame.

2) Lowercasing: All text was converted to lowercase to maintain consistency and minimize variability due to case sensitivity, using the lower() function.

3) Removal of URLs and mentions: Hyperlinks and user mentions, which do not contribute meaningful information to the sentiment classification task, were eliminated through regular expressions.

4) Removal of punctuation and digits: Punctuation marks and numerical digits were removed using standard string processing techniques to focus solely on the textual content relevant for semantic analysis.

5) Lemmatization: Lemmatization was applied to normalize words to their base or dictionary forms by utilizing vocabulary and morphological analysis. This step helps in reducing inflectional forms and improving the semantic understanding of the text.

After completing the text cleaning procedures and encoding the sentiment labels, the dataset was prepared for further processing. Table III presents a sample of the cleaned text alongside the corresponding encoded sentiment labels.

TABLE III. SAMPLE OF CLEANED TEXT AND CORRESPONDING ENCODED SENTIMENT LABELS AFTER PREPROCESSING

Index	Cleaned Text	Encoded Sentiment
0	algebra math game saying create game incorpora...	1
1	peer review module fully done anything wrong p...	1
2	grow brain right middle front room statement f...	1
3	math right wrong math become conceptual adapt ...	1
4	district group group based struggling idea tim...	1
...	...	...
29599	dear option regular best josh	2
29600	fabulous typo module slide title supposed viol...	2
29601	thanks josh hint anon screen name.	2
29602	whoa nut thanks value calculator	2
29603	thanks	2

The cleaned text will serve as input for the subsequent tokenization, encoding, and embedding processes, while the encoded sentiment labels will be used as target outputs during supervised model training.

D. Data Visualization

Data visualizations are an essential aspect of exploring and understanding datasets. In this study, we employed visual techniques such as word clouds and word frequency analysis to gain insights into the sentiment distribution and the characteristics of learner posts. These visualizations help to identify underlying patterns and potential imbalances within the dataset, providing valuable context for the sentiment analysis task. The following sections will elaborate on the key visualizations utilized in this study and their role in uncovering meaningful trends within the data.

1) Word cloud: The word cloud visualization highlights the most frequently occurring words within a dataset. Words that appear more often are displayed in larger fonts, while those used less frequently are shown in smaller fonts. Fig. 7 presents a word cloud that provides an overview of the emotional trends expressed in the posts, encompassing positive, negative, and neutral sentiment words in a single, comprehensive visualization. This allows for a clear understanding of the language patterns within the dataset.

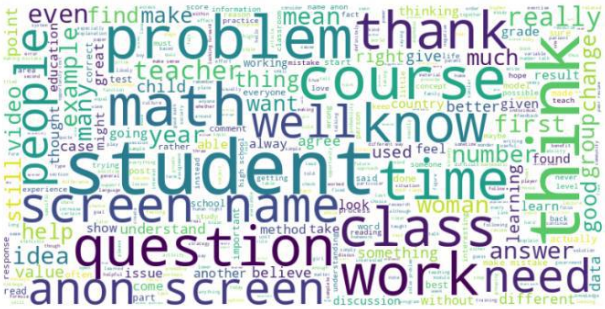

Fig. 7. Word cloud.

2) Words frequency: Word frequency analysis provides essential insights into the language used within a text or corpus. This analysis allows for the identification of recurring terms and key phrases, revealing patterns and underlying themes within the dataset. Fig. 8, 9 and 10 illustrate the most frequently occurring words associated with each sentiment category. In positive sentiment posts, terms such as "great", "learning," and "thanks" are prominent, reflecting positive engagement and appreciation for the course. Conversely, negative sentiment posts highlight words like "problem", "teacher" and "grade", indicating issues or dissatisfaction encountered by stude,nts. Neutral sentiment posts feature terms like "question", "answer", and "data", commonly found in objective discussions about course content without significant emotional emphasis.

E. Data Augmentaiton

Text data augmentation is a technique in natural language processing (NLP) that expands the size and diversity of a text dataset by generating variations of existing data. By introducing these variations, models become more robust and generalizable, improving their performance on new, unseen data. Data augmentation helps reduce overfitting, ensuring that models can handle diverse and unpredictable inputs in real-world scenarios.

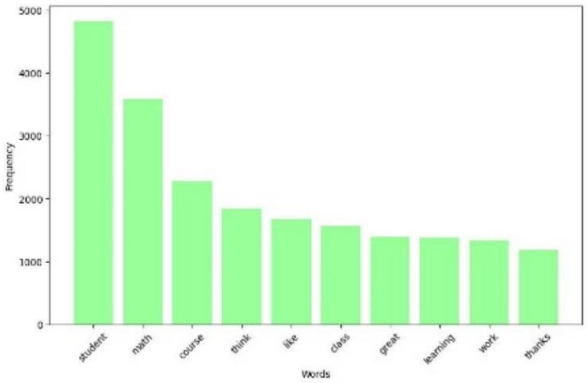

Fig. 8. Top 10 most frequent positive words.

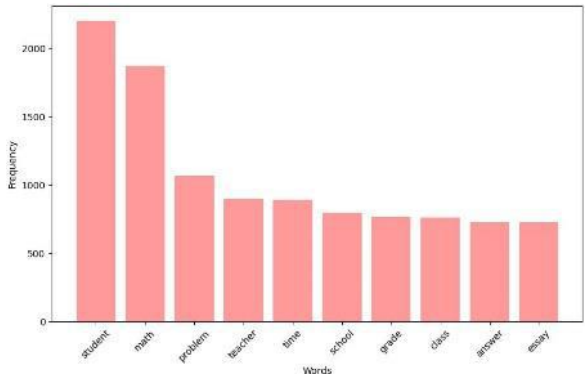

Fig. 9. Top 10 most frequent negative words.

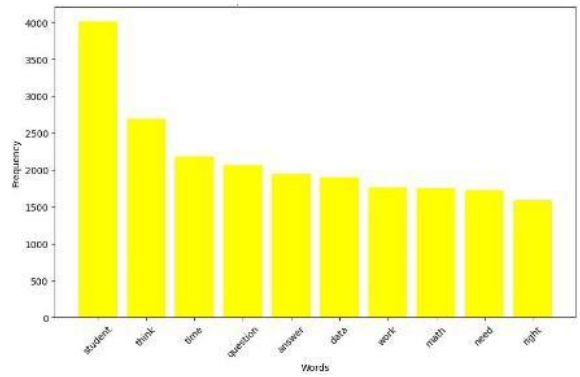

Fig. 10. Top 10 most frequent neutral words.

In this study, we applied the following data augmentation techniques:

1) Synonym replacement: Words are replaced with their synonyms to introduce variation without altering the overall meaning.

2) Random insertion: Random words are inserted into the text to diversify the vocabulary and sentence structure.

3) Random swap: The positions of random words in the text are swapped to generate different syntactical structures.

4) Random deletion: Random words are removed from the text to simulate missing information and prevent overfitting.

5) Character-level augmentation: The text is modified at the character level, such as introducing typos, to simulate real-world text input errors.

Each of these methods generates variations of the original text, thus increasing the diversity of the dataset and enhancing its representativeness. By applying these techniques, we aim to improve the model's ability to generalize and reduce overfitting.

F. Text Representation

In order to prepare the textual data for input into the BiLSTM model, it is necessary to transform raw text into a structured numerical format that preserves both semantic and contextual information. This transformation comprises three key steps: tokenization, encoding, and embedding. Tokenization decomposes the text into individual units (tokens) suitable for computational processing. Encoding subsequently maps these tokens into unique integer identifiers, forming standardized input sequences. Finally, embedding projects these encoded sequences into dense vector spaces that capture semantic relationships between words. The following sub-sections elaborate on each of these steps.

1) Tokenization: Tokenization was performed using the Tokenizer class from the TensorFlow Keras library. The Tokenizer constructs a vocabulary from the text corpus and converts the textual data into sequences of integers suitable for input to the BiLSTM model. An instance of the Tokenizer was initialized to process the corpus, creating an empty dictionary to map each unique word to a distinct integer index. This mapping establishes the basis for subsequent encoding and embedding procedures.

2) Encoding: After the text had been divided into tokens, each token was assigned a unique integer identifier based on a constructed vocabulary. This encoding process transformed the sequences of tokens into sequences of integers, enabling standardized numerical input for the BiLSTM model. Moreover, the target variable, representing the sentiment class (e.g., positive, neutral, negative), was also encoded numerically to facilitate the supervised learning process. By ensuring that both the input features and the output labels were numerically represented, the data became suitable for effective computational modeling and training.

3) Word embedding: After encoding, the integer sequences were transformed into dense vector spaces through the use of pre-trained word embeddings. Word embeddings capture the semantic and syntactic properties of words, enabling the model to leverage semantic relationships for improved predictive performance.

The embedding layer maps each token index to its corresponding vector representation, effectively addressing issues of data sparsity and reducing the number of trainable parameters, which in turn mitigates the risk of overfitting. In this work, pre-trained word vectors from GloVe (Global Vectors for Word Representation), an unsupervised learning algorithm introduced by Stanford researchers in 2014 [33], were utilized. These vectors were employed to initialize the embedding layer, with each word's embedding serving as the initial weight in the model. This initialization enables faster convergence and more effective learning during model training.

G. Model Architecture

The architecture of the proposed model was designed to effectively capture semantic and contextual features from learners' posts for sentiment classification. It is based on a Bidirectional Long Short-Term Memory (BiLSTM) deep neural network, augmented with several regularization techniques to enhance generalization performance.

The model begins with an embedding layer, which converts input tokens into dense vectors of a fixed size. This layer was initialized with pre-trained GloVe vectors to embed semantic information into the input representations. To prevent overfitting at the embedding level, a SpatialDropout1D layer was applied, which randomly drops entire 1D feature maps to promote robust feature learning.

Following the embedding and dropout operations, two stacked Bidirectional LSTM layers were employed. The first BiLSTM layer consists of 128 units, processes the input sequences in both forward and backward directions, and applies both dropout and recurrent dropout for regularization. Batch normalization was applied after this layer to stabilize and accelerate the training process. The second BiLSTM layer, consisting of 64 units, further refines the sequential features using a similar configuration of dropout, recurrent dropout, and batch normalization.

After the stacked BiLSTM layers, the model includes a fully connected dense layer with 64 units and ReLU activation, introducing non-linearity to capture more complex patterns within the extracted features. A standard dropout layer was subsequently added to provide further regularization and reduce the risk of overfitting.

Finally, the model concludes with a dense output layer utilizing a softmax activation function, producing probabilistic outputs across the sentiment classes. This enables the model to perform multi-class sentiment classification by assigning a probability score to each class.

Overall, this architecture effectively balances the need to capture intricate sequential dependencies with robust regularization mechanisms, resulting in a model that generalizes well to unseen data. The overall architecture of the proposed model is illustrated in Fig. 11.

H. Training Procedure

The dataset was split into three parts: 60% for training, 10% for validation, and 30% for testing. The training data (60%) was used to build and train the model, while the validation data (10%) helped tune the model during training, and the testing data (30%) was reserved for final evaluation. Padding was applied to both the training and testing datasets to ensure uniform sequence lengths for efficient batch processing. The model was trained with a batch size of 64 for 50 epochs using the Adam optimizer with a learning rate of 0.001. Categorical crossentropy was used as the loss function for multi-class classification.

Table IV summarizes the training hyperparameters and their justifications.

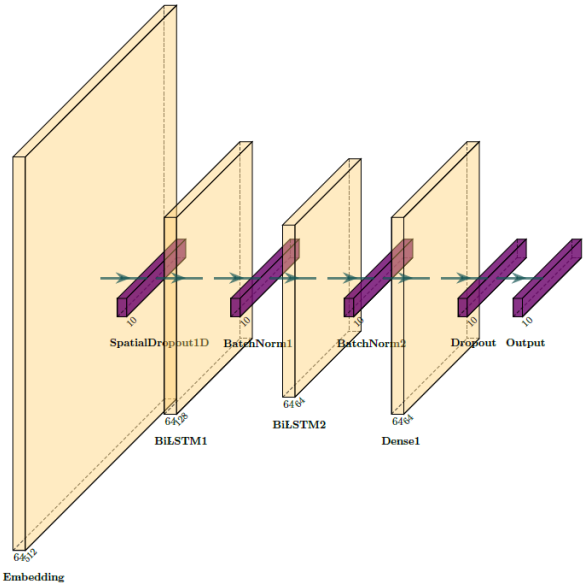

Fig. 11. Model architecture.

TABLE IV. TRAINING HYPERPARAMETERS AND THEIR JUSTIFICATIONS

Parameter	Details	Justification
Dataset Split	**60%** Training, **10%** Validation, **30%** Testing	Ensures fair model evaluation and prevents data leakage.
Batch Size	64	Balances computational efficiency and model stability.
Epochs	50	Allows sufficient training while preventing overfitting.
Optimizer	Adam	Accelerates convergence and adapts learning rates.
Loss Function	Categorical Crossentropy	Suitable for multi-class classification problems.
Early Stopping	Enabled (based on validation loss)	Prevents overfitting by stopping training when performance plateaus

IV. RESULTS

In this section, we present the experimental results and a comparative analysis of the performance of our proposed enhanced Bi-LSTM model against several established baseline algorithms. The evaluation focuses on key metrics, including accuracy, precision, recall, and F1-score, to comprehensively assess the effectiveness and efficiency of the approach. In addition to presenting the numerical results, we provide detailed interpretations and discussions to highlight the significance of the findings, compare them with related works, and address the strengths and limitations of the model within the MOOC sentiment analysis context.

A. Evaluation Metrics

To comprehensively evaluate the performance of the proposed sentiment analysis model, several widely used classification metrics were employed, including accuracy, precision, recall, and F1-score. These metrics provide a robust understanding of the model's effectiveness across different aspects of sentiment classification, beyond mere accuracy alone. Their definitions and corresponding formulas are as follows:

- Accuracy: measures the fraction of predictions where the model made a correct decision. It is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

- Precision: is the ratio of true positive results to all predicted positive results. It is calculated as:

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

- Recall: is the ratio of true positive results to all actual positive samples. It is computed as:

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

- F1-score: represents the harmonic mean between precision and recall, providing a balanced evaluation metric. It is expressed as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (10)$$

Where TP denotes True Positives, TN denotes True Negatives, FP denotes False Positives, and FN denotes False Negatives.

B. Results and Comparison with Baseline Models

In this section, we present the experimental results of the proposed enhanced Bi-LSTM model on the Stanford MOOC Posts dataset and compare its performance against several baseline machine learning models, including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Multilayer Perceptron (MLP).

The evaluation was carried out using the previously described metrics: accuracy, precision, recall, and F1-score. Furthermore, the experiments were conducted under two conditions: without data augmentation and with data augmentation, to assess the impact of augmentation techniques on model performance.

To ensure a comprehensive internal evaluation, we reported multiple evaluation metrics, including Accuracy, Precision, Recall, and F1-score. Considering the moderate class imbalance present in the Stanford MOOC Posts dataset, particularly the predominance of neutral sentiment posts the F1-score was particularly informative for assessing balanced classification performance beyond what Accuracy alone could capture.

The detailed results for each model under both conditions are summarized in Table V.

TABLE V. EXPERIMENT RESULTS

	No Data augmentation				**Data augmentation**			
	Acc	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**
SVM	71.79	71.37	71.80	69.82	91.32	91.57	91.27	91.23
DT	61.56	60.94	61.57	61.21	82.74	82.60	82.90	82.59
RF	70.40	69.71	70.38	67.89	90.86	91.85	90.85	90.77
LR	71.39	70.52	71.38	70.29	71.03	70.27	71.05	69.87
MLP	67.59	66.81	67.50	67.04	89.77	89.81	89.79	89.70
BI-LSTM	**71.22**	**71.19**	**70.98**	**71.13**	**88.54**	**88.51**	**88.55**	**88.52**

After observing the experimental results, it is evident that the Bi-LSTM model achieved strong performance across all evaluation metrics. Without data augmentation, the Bi-LSTM achieved an accuracy of 71.22% and an F1-score of 71.13%, outperforming baseline models such as SVM (69.82% F1-score) and RF (67.89% F1-score). After applying data augmentation techniques, the Bi-LSTM's accuracy increased to 88.54%, with an F1-score of 88.52%. These results confirm the model's effectiveness in both overall prediction correctness (Accuracy) and balanced classification performance (F1-score).

Although the SVM model achieved the highest F1-score of 91.23% after augmentation, the Bi-LSTM demonstrated consistent and competitive performance across all evaluation metrics. The slight and unexpected outperformance of SVM can be attributed to the structured nature of the dataset, where traditional machine learning models can sometimes perform better in recognizing more formal, less noisy textual patterns. Nevertheless, the Bi-LSTM model showed strong robustness and generalization capabilities, particularly when considering its potential scalability to larger and more diverse datasets with higher linguistic variability.

The significant improvement observed after applying data augmentation techniques can be attributed to the increased diversity and richness of the training data. By generating synthetic examples through operations such as synonym replacement, random insertion, word swapping, and random deletion, the model was exposed to a wider variety of linguistic patterns and textual variations. This exposure helped the Bi-LSTM model generalize better to unseen data, reduce overfitting, and become more robust in handling informal expressions, typos, and abbreviations commonly found in learner-generated posts. Consequently, the augmented dataset enabled the model to capture the underlying sentiment signals more effectively, leading to notable gains across all evaluation metrics.

C. Comparison with Existing Studies

After evaluating the performance of the proposed enhanced Bi-LSTM model internally, this section presents a comparative analysis against previously published sentiment analysis approaches that also used the Stanford MOOC Posts dataset. Accuracy is used as the primary evaluation metric to allow a consistent and meaningful comparison with results reported in existing studies. The comparative results are summarized in Table VI.

After reviewing the comparative results presented in Table VI, it is evident that the proposed enhanced BiLSTM model achieves a highly competitive performance, attaining an accuracy of 88.54%. This surpasses the results of several existing approaches, including the HAN-based method by Chanaa and El Faddouli [34] (70.3%), the XLNet-CNN model by Farahmand et al. [35] (77%), and the LSTM-based framework by Munigadiapa and Adilakshmi [36] (87.64%). Although the SSDL approach proposed by Chen et al. [10] achieved a slightly higher accuracy of 89.73%, it relies on a semi-supervised learning strategy and the integration of multiple embeddings, increasing the model's complexity. In contrast, the proposed BiLSTM model demonstrates strong performance using a simpler architecture enhanced with GloVe embeddings and data augmentation techniques, making it a more practical and efficient solution for large-scale MOOC sentiment analysis tasks.

TABLE VI. COMPARATIVE RESULTS

Study	Techniques Applied	Accuracy	Comments
A. Chanaa and N. El Faddouli [34]	HAN	70.3%	Utilizes a Hierarchical Attention Network (HAN) to surpass traditional text classification models.
J. Chen, J. Feng, X. Sun, and Y. Liu [10]	SSDL	89.73%	Proposes a co-training semi-supervised deep learning framework (SSDL) that combines word embedding and character-based embedding to improve sentiment classification.
Farahmand et al.[35]	XLNet-CNN	77%	Identifies and visualizes student sentiment in discussion forums to enhance self-awareness and engagement, with sentiments categorized as negative, neutral, or positive.
Munigadiapa, P., Adilakshmi, T. [36]	LSTM, GloVe embedding, Ax	87.64%	Proposes a sentiment analysis system using a new LSTM architecture and Ax hyperparameter tuning, designed for large-scale sequential sentiment analysis.
Our Study	BiLSTM, GloVe,embedding	**88,54%**	Proposes an enhanced BiLSTM model utilizing GloVe embeddings and data augmentation techniques to improve sentiment classification performance.

V. DISCUSSION

The experimental results demonstrate the effectiveness of the proposed enhanced Bi-LSTM model for sentiment analysis in the MOOC context. After applying data augmentation techniques, the Bi-LSTM model exhibited substantial improvements across all evaluation metrics, confirming the benefits of enriching the training data to better capture the linguistic variability present in learner-generated posts [37].

An interesting observation was the slight and unexpected outperformance of the SVM model in terms of F1-score after data augmentation. This deviation highlights that, in relatively structured and less noisy datasets, traditional machine learning models can sometimes capitalize on clear textual patterns more efficiently than deep neural networks [38], which typically require larger and more heterogeneous datasets to fully realize their advantages. Nevertheless, the Bi-LSTM model demonstrated strong generalization capabilities across all evaluation metrics, particularly in terms of its scalability to more complex and diverse data environments.

When compared with existing studies on MOOC sentiment analysis that used the same Stanford MOOC Posts dataset, the proposed Bi-LSTM framework achieved competitive performance. Although some prior works reported slightly

higher accuracy scores, the present study emphasizes robustness, stability, and real-world applicability across diverse sentiment categories. The integration of data augmentation and regularization strategies proved essential in enhancing the model's ability to generalize, aligning with broader trends observed in recent natural language processing research. Given these generalization capabilities, the framework may also be adaptable to other domains involving informal or user-generated content, such as product reviews, social media streams, customer sentiment analysis or hate speech detection [39], where similar linguistic variability and class imbalance are present.

The proposed approach is characterized by several strengths, including the ability to handle class imbalance, improve performance on noisy text, and adapt to evolving online discourse. Nonetheless, certain opportunities remain for further extension. While the data augmentation techniques employed in this study, such as synonym replacement and random word swapping, proved highly effective, future work could investigate complementary strategies such as contextual augmentation using masked language models (e.g. BERT-based augmentation) or back-translation to further diversify the training set. Additionally, building upon the demonstrated effectiveness of the enhanced Bi-LSTM model, future research could explore the integration of transformer-based architectures such as BERT [40], RoBERTa [41], or ALBERT [42] to capture even deeper contextual relationships and subtle semantic nuances present in learner-generated posts, thereby expanding the model's capabilities for more complex and dynamic educational environments.

VI. Conclusion

This study proposed an enhanced Bi-LSTM framework for sentiment analysis of learners' posts within the MOOC context. By integrating carefully designed data preprocessing, data augmentation techniques, and regularization strategies, the model demonstrated robust performance across multiple evaluation metrics. Experimental results confirmed the effectiveness of the proposed approach, with notable improvements in both Accuracy and F1-score after applying data augmentation, highlighting the model's ability to generalize across varied learner-generated content.

A comparative analysis with traditional machine learning models, including SVM, Decision Tree, Random Forest, Logistic Regression, and MLP, showed that the enhanced Bi-LSTM model achieved competitive results, particularly in balancing precision, recall, and F1-score. Although a slight and unexpected outperformance by SVM was observed under specific conditions, the Bi-LSTM model consistently demonstrated strong adaptability and scalability, positioning it as a promising solution for sentiment analysis tasks in large-scale educational environments.

The findings of this study contribute to advancing the field of educational sentiment analysis by providing a scalable and robust framework capable of addressing real-world challenges such as informal language, typographical errors, class imbalance, and varied textual structures. In particular, the integration of data augmentation introduced valuable linguistic variations, reduced overfitting, and contributed to enhancing the model's ability to generalize across diverse linguistic patterns present in learner-generated content. These contributions support the development of more adaptive, sentiment-aware learning support systems, benefiting researchers and practitioners aiming to improve learner engagement and personalized feedback in online education.

For future work, several promising directions are identified. Extending the framework to multilingual datasets would enable broader applicability across diverse learning environments and cultural contexts. Furthermore, incorporating more sophisticated augmentation strategies, such as syntax-aware or semantics-driven transformations, could further enrich the training data. Additionally, extending the proposed model to related NLP tasks, such as emotion detection or sarcasm analysis, could leverage its ability to capture nuanced contextual relationships, offering further valuable applications. The integration of the proposed framework into real-world educational support systems represents a valuable next step, enabling instructors to monitor learner sentiment in real time and tailor instructional strategies to enhance engagement. Finally, exploring transformer-based architectures, such as BERT or RoBERTa fine-tuned for educational sentiment analysis, also holds potential to enhance classification performance and advance sentiment analysis capabilities in online learning platforms.

References

[1] A. Kastrati, A. Imran, and B. Kastrati, "Weakly supervised framework for aspect-based sentiment analysis on MOOC comments," Computers and Education: Artificial Intelligence, vol. 2, 2021, doi: 10.1016/j.caeai.2021.100020.

[2] D. Zhang and Y. Zhu, "Hybrid attention-based neural networks for sentiment and emotion classification in education-related texts," Applied Sciences, vol. 12, no. 10, 2022, doi: 10.3390/app12104784.

[3] H. T. T. Phan, T. T. Nguyen, T. T. H. Nguyen, and A. V. Nguyen, "Aspect-based sentiment analysis in education using hybrid deep learning," International Journal of Advanced Computer Science and Applications, vol. 15, no. 1, 2024, doi: 10.14569/IJACSA.2024.0150172.

[4] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," Computers in Human Behavior, vol. 31, pp. 527–541, 2014, doi: 10.1016/j.chb.2013.05.024.

[5] 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.22202.

[6] R. Ramesh, D. Y. Huang, and A. C. Kok, "Predicting student dropout in MOOCs using deep learning with attention mechanism," Education and Information Technologies, vol. 28, pp. 8791–8810, 2023, doi: 10.1007/s10639-023-11786-2.

[7] 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.

[8] R. Alatrash, H. Ezaldeen, R. Misra, and R. Priyadarshini, "Sentiment analysis using deep learning for recommendation in e-learning domain," Progress in Advanced Computing and Intelligent Engineering, Springer, pp. 123–133, 2021, doi: 10.1007/978-981-15-4032-5_12.

[9] Y. Zhang and Y. Zhu, "Sentiment-content analysis of user reviews in LMOOCs," Interactive Learning Environments, vol. 30, no. 1, pp. 134–150, 2022, doi: 10.1080/10494820.2021.1908277.

[10] J. Chen, J. Feng, X. Sun, and Y. Liu, "Co-training semi-supervised deep learning for sentiment classification of MOOC forum posts," Symmetry, vol. 12, no. 1, p. 8, 2019, doi: 10.3390/sym12010008.

[11] K. Sailunaz and R. Alhajj, “Emotion and sentiment analysis from Twitter text,” Journal of Computational Science, vol. 36, p. 101003, 2019, doi: 10.1016/j.jocs.2019.05.009.

[12] A. Kumar, A. Ekbal, D. Kawahra, and S. Kurohashi, “Emotion helps sentiment: A multi-task model for sentiment and emotion analysis,” IEEE Transactions on Affective Computing, vol. 14, no. 1, pp. 318–330, 2023, doi: 10.1109/TAFFC.2020.2996911.

[13] Y. Zhang, H. Xu, and L. Zhang, “Attention-based LSTM for aspect-level sentiment classification,” Cognitive Computation, vol. 14, pp. 1235–1246, 2022, doi: 10.1007/s12559-022-10025-5.

[14] R. Priyadharshini, V. Vaidehi, P. S. Kumar, and M. Janakiraman, “A hybrid deep learning approach for sentiment analysis using CNN and Bi-LSTM,” International Journal of Intelligent Engineering and Systems, vol. 14, no. 6, pp. 181–190, 2021, doi: 10.22266/ijies2021.1231.17.

[15] A. S. Hornby, Oxford Advanced Learner’s Dictionary. Emotion, Oxford University Press, 2000.

[16] K. R. Scherer, “What are emotions? and how can they be measured?,” Social Sciences Information, vol. 44, no. 4, pp. 695–729, 2005, doi: 10.1177/0539018405058216.

[17] S. Jain and K. Asawa, "Modeling of emotion elicitation conditions for a cognitive-emotive architecture," Cognitive Systems Research, vol. 52, pp. 535–548, Dec. 2018, doi: 10.1016/j.cogsys.2018.12.012.

[18] C. Darwin, The Expression of the Emotions in Man and Animals, University of Chicago Press, 2015.

[19] W. B. Cannon, “The James-Lange theory of emotions: A critical examination and an alternative theory,” The American Journal of Psychology, vol. 39, no. 1/4, pp. 106–124, 1927, doi: 10.2307/1415404.

[20] S. Schachter and J. Singer, “Cognitive, social, and physiological determinants of emotional state,” Psychological Review, vol. 69, no. 5, pp. 379–399, 1962, doi: 10.1037/h0046234.

[21] J. A. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.

[22] P. Ekman, “An argument for basic emotions,” Cognition and Emotion, vol. 6, no. 3–4, pp. 169–200, 1992, doi: 10.1080/02699939208411068.

[23] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, “New avenues in opinion mining and sentiment analysis,” IEEE Intelligent Systems, vol. 28, no. 2, pp. 15–21, 2013, doi: 10.1109/MIS.2013.30.

[24] B. Liu, “Sentiment analysis and opinion mining,” Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[25] W. X. Zhao et al., “Topical keyphrase extraction from Twitter,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 7, pp. 1852–1865, Jul. 2016, doi: 10.1109/TKDE.2016.2535384.

[26] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, “Sentiment analysis is a big suitcase,” IEEE Intelligent Systems, vol. 32, no. 6, pp. 74–80, 2017, doi: 10.1109/MIS.2017.4531228.

[27] Graves, A.: Long short-term memory. In: Supervised Sequence Labelling with Recurrent Neural Networks, pp. 37–45. Springer, Berlin, (2012)

[28] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” Neural Networks, vol. 18, no. 5-6, pp. 602–610, 2005.

[30] M. Yang, S. Lee, J. Choi, and H. Kim, "A Bidirectional LSTM Language Model for Code Evaluation and Repair," Symmetry, vol. 13, no. 2, p. 247, 2021, doi: 10.3390/sym13020247.

[31] H. Kim, J. Jeong, and H. Kim, "Bi-LSTM Model to Increase Accuracy in Text Classification," Applied Sciences, vol. 10, no. 17, p. 5841, 2020, doi: 10.3390/app10175841.

[32] A. Agrawal and A. Paepcke. (2014).The Stanford MOOCPosts Data Set. Accessed:May.18,2024.[Online].Available:https://datastage.stanford.edu/StanfordMoocPosts/

[33] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1532–1543, 2014.

[34] A. Chanaa and N. El Faddouli, “E-learning text sentiment classification using hierarchical attention network (HAN),” International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 13, p. 157, Jul. 2021, doi: 10.3991/ijet.v16i13.2257

[35] Farahmand, A., Dewan, M.A.A., Lin, F., Hwang, WY. (2023). Improving Students’ Self-awareness by Analyzing Course Discussion Forum Data. In: Sottilare, R.A., Schwarz, J. (eds) Adaptive Instructional Systems. HCII 2023. Lecture Notes in Computer Science, vol 14044. Springer, Cham. https://doi.org/10.1007/978-3-031-34735-1_1

[36] Munigadiapa, P. and Adilakshmi, T., 2023. MOOC-LSTM: The LSTM Architecture for Sentiment Analysis on MOOCs Forum Posts. In: R. Buyya, S.M. Hernandez, R.M.R. Kovvur and T.H. Sarma, eds. Computational Intelligence and Data Analytics. Singapore: Springer

[37] J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 6383–6389, 2019. doi: 10.18653/v1/D19-1670.

[38] H. Fu, H. Song, W. Cui, and L. Wang, “A comparative study of deep learning performance on sentiment classification tasks,” IEEE Access, vol. 8, pp. 94960–94971, 2020. doi: 10.1109/ACCESS.2020.2994969.

[39] B. Kumar, R. Verma, and A. K. Sharma, “MLHS-CGCapNet: A Lightweight Model for Multilingual Hate Speech Detection,” IEEE Access, vol. 12, pp. 12345–12360, 2024. doi: 10.1109/ACCESS.2024.3434664

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186. doi: 10.48550/arXiv.1810.04805.

[41] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *arXiv preprint*, arXiv:1907.11692, 2019. doi: 10.48550/arXiv.1907.11692.

[42] Z. Lan et al., “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,” in *Proc. ICLR*, 2020. Available: https://arxiv.org/abs/1909.11942.