Enhancing Urgent Post Classification in MOOC Discussion Forums: A Fusion of Neural Network with BERT Embedding

Mujtaba Sultani (mujtaba.sultani@sru.ac.ir**), Negin Deneshpour*(**ndaneshpour@sru.ac.ir**)**
Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran

ABSTRACT

MOOCs serves as platforms where students engage with instructors and voice educational concerns. However, the challenge of distinguishing urgent posts among a large volume necessitates automated classification. As education evolves, addressing high dropout rates and respond to urgent learner queries becomes more important. This research focuses on binary classification of student posts in the Stanford MOOC Posts dataset into urgent and not urgent categories. Previous studies employed various techniques, such as traditional machine learning algorithms, deep learning, and embeddings like BERT, to address this classification challenge. However, there remains room for enhancement in the identification of urgent posts, particularly in handling imbalanced datasets. Our model, named CBiLSTM, is a fusion of CNN and BiLSTM layers, enhanced by BERT-based contextual embeddings. To address class imbalance, BERT-based data augmentation is applied. The evaluation involves comparison with baseline models, including standalone CNN, LSTM, and BiLSTM, as well as state-of-the-art models. The proposed model boasted a 93.3% F1-weighted score and 84.1% F1 score for the urgent class. This represents a notable improvement of 0.6% and 0.8% over the best-performing state-of-the-art model, respectively.

INDEX TERMS

BERT, Data Augmentation, CBiLSTM, Urgent post classification, MOOC

* Corresponding author.
E-mail address: ndaneshpour@sru.ac.ir(N. Daneshpour).

1. INTRODUCTION

Educational technology has been notably shaped by the growing prevalence of Massive Online Open Courses (MOOCs). These courses aim to make education accessible worldwide and provide access to online resources (Zhenghao, Alcorn, Christensen, Eriksson, Koller, & Emanuel, 2015). The influence of MOOCs can be seen in the numbers: More than 58 million users worldwide have enrolled in a MOOC. Currently, more than 220 million learners are participating in MOOCs worldwide (A Decade of MOOCs, 2021). The widespread popularity of MOOCs is also evident from the fact that more than 700 prestigious universities participate in them and offer a variety of courses accessible through platforms such as Coursera, edX, and Udacity. The wide availability of MOOCs signifies a shift in educational approaches that emphasizes the democratization of knowledge and promotes a culture of continuous, lifelong learning (Shah, 2016). Individuals can use these courses to learn new skills, enhance their understanding, and foster professional development according to their preferences (Zhenghao et al., 2015).

A key aspect of MOOCs is the inclusion of communication platforms, particularly discussion forums. These forums facilitate interaction between learners and instructors, as well as peer-to-peer exchanges (Zhang, Chen, & Phang, 2018). These forums play an important role in supporting different learning approaches that shape the educational experiences of MOOC participants. Moreover, these forums serve as a valuable channel for students to articulate their questions and urgent concerns (Feng, Chen, Zhao, Chen, & Xi, 2015). However, considering the substantial quantity of MOOC participants and the limited number of instructors, it poses a challenge to effectively track and respond to students' posts and questions. Quick responses to important posts are often necessary to help students overcome obstacles during their learning journey. Failure to provide timely feedback can lead to learner frustration and increase dropout rates (Hone & El Said, 2016). Therefore, it is necessary to develop mechanisms that differentiate urgent posts and ensure that they receive immediate attention and response from instructors (Almatrafi, Johri, & Rangwala, 2018). Implementing an effective system for monitoring and handling urgent posts would allow instructors to give precedence to their responses and effectively handle the overwhelming volume of submissions. This system would not only optimize instructors time and attention, but also empower them to devote more energy to promote community engagement and provide valuable support (Almatrafi et al., 2018).

On the other hand, model generalization, which denoting the ability of a machine learning model to respond accurately to unseen data, is a critical aspect of model development (Birjali, Kasri, & Beni-Hssane, 2021). However, the presence of a generalization constraint, characterized by a significant performance difference between training data and new, unseen data, poses a significant challenge when applying machine learning models, especially in subdomains or specific contexts where training data may be limited or unavailable. In the education domain, effective model generalization is of immense importance. Educational data are inherently diverse, encompassing different subjects, learning styles, and student demographics. Therefore, models must have the ability to adapt and perform reliably in different subdomains of the education domain (Birjali et al., 2021).

By effectively closing the generalization gap, machine learning models can provide more accurate and reliable insights, recommendations, and support for learners, educators, and educational institutions. Closing this limitation contributes to the creation of personalized and tailored learning experiences, improved learning outcomes, and more effective educational interventions (Birjali et al., 2021).

Extensive research has focused on classifying MOOC forum posts into urgent and non-urgent categories. Various word representation techniques and classification approaches have been explored to develop effective models. The goal is to prioritize posts according to their urgency. In (Almatrafi et al., 2018, Agrawal, Venkatraman, Leonard, & Paepcke, 2015, Chang, Lee, Wu, Liu, & Liu, 2021), statistical techniques like term frequency (TF), inverse document frequency (IDF), and term frequency-inverse document frequency (TF-IDF) were employed to represent words. However, they neglected the meaning of word order (Xue, & Chen, 2022), which led to a limitation in capturing the document context. These studies used conventional classification algorithms like SVM and Nearest Centroid, which require low computational effort but have poor performance because they often rely heavily on manual feature selection (El-Rashidy, Farouk, El-Fishawy, Aslan, & Khodeir, 2023).

Studies (El-Rashidy et al., 2023; Sun, Guo, Gao, Zhang, Xiao, & Feng, 2019; Guo, Sun, Wang, Gao, & Feng, 2019; Khodeir, 2021) have used pre-trained models like Google News and Glove to represent words with dense vectors that capture their contextual meaning within the document (Almeida, & Xexéo, 2019). These studies used different architectures including multiple CNNs, CNN aggregation, GRU, and attention layers to develop their models. They emphasized on including additional representational features, choosing effective features, and assigning higher weight to the most important features. However, none of these studies addressed the problem of imbalanced datasets, which can affect classification performance and lead to bias toward larger classes (Wei, & Zou, 2019; Guo, Yin, Dong, Yang, & Zhou, 2008).

Improvements in the classification of imbalanced datasets have been classified into five categories: Data, Algorithms, Cost-Sensitive, Feature Selection, and Ensemble Approaches (Ramyachitra, & Manikandan, 2014). In this study, we specifically address the data level by focusing on techniques for balancing text datasets. Common data-level methods include re-sampling to adjust the number of samples in the dataset. Oversampling involves adding samples, usually by copying samples, while under-sampling involves removing samples, often by random selection. While these methods have shown some effectiveness in data matching, they are not sufficient to completely solve the problem at hand (Branco, Torgo, & Ribeiro, 2016).

We employed BERT -based data augmentation (DA) to solve the data balancing problem (Kumar, Choudhary, & Cho, 2020). Data augmentation stands as a commonly employed method for increasing the scale of training data. This approach significantly contributes in mitigating overfitting and improving the robustness of machine learning models, especially for tasks with limited data availability (Kumar et al., 2020). Moreover, we employed BERT model as the embedding layer, fine-tuned by a novel hybrid deep learning approach, combining Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) layers. We named this hybrid model Convolutional BiLSTM (CBiLSTM), which enables an effective and precise classification of urgent posts. In

CBiLSTM, the CNN layers effectively handle the high dimensionality of input texts, and BiLSTM layers explore feature context bidirectionally. The outcomes illustrate that CBiLSTM outperforms conventional deep learning models commonly utilized for urgent post classification using widely recognized Stanford MOOCPost Corpus defined in (Almatrafi et al., 2018). Key contributions of this paper include:

1. To tackle with data imbalance, we employ BERT-based data augmentation technique to balance the training dataset. This technique significantly enhances the model's performance and generalization, crucial for accurate classification in imbalanced scenarios. Our study is the first to employ BERT-based data augmentation for dataset balancing purpose in MOOC context.
2. To leverage the power of contextualized embeddings, we adopt the BERT pre-trained model as an embedding layer. By utilizing BERT's language understanding capabilities, our model gains deeper insight into the context and semantic meaning of the input text, resulting in more accurate and context-aware representations.
3. Proposing CBiLSTM, a hybrid model that integrates the power of CNN and BiLSTM. This fusion enables our model to extract local features with CNN by analyzing spatial relationships within the data, while effectively capturing long-term dependencies with BiLSTM. The combination of these two architectures allows our model to produce more accurate and robust predictions.
4. Our approach achieves better results for the classification of urgent posts in MOOCs using the Stanford MOOC Posts dataset, surpassing existing methods.

The following sections of this paper provide a structured flow of our research. Section 2 offers a brief overview of related work, while Section 3 delves into the adapted method. In Section 4, we detail the experimental setup, data, and evaluation metrics. Results are discussed in Section 5, followed by Section 6. Finally, Sections 7 and 8 cover future work and conclusive results.

2. RELATED WORK

In the field of online education, MOOCs have become the most popular choice as they provide students from all walks of life with a flexible and accessible learning experience. With the ever-growing number of participants in MOOCs, discussion forums have become important centers for communication, knowledge sharing, and collaborative learning (Liyanagunawardena, Adams, & Williams, 2013). Nevertheless, the abundance of student contributions in these forums poses a significant challenge for instructors to respond on time, especially when urgent questions or concerns arise.

Consequently, researchers have focused to the classification of urgent posts within MOOC discussion forums to enhance the effectiveness and efficiency of instructor-learner interactions. Here we aim to review existing research in this area and examine various methods and approaches for classifying and prioritizing urgent posts in MOOC environments. MOOC post classification employs both traditional machine learning and deep learning algorithms (Guo et al., 2019).

In their study, (Almatrafi et al., 2018) employed metadata, linguistic features, and traditional machine learning algorithms to identify urgent posts in MOOCs. AdaBoost classification algorithm provided the better results. (Bakharia, 2016) developed a comprehensive classification model considering the dimensions of urgency, sentiment, and confusion in different domains. However, this work is still in its early stages and is not sufficiently generalizable in classifying urgent posts because it gives good results within specific courses or domains but is difficult to generalize across different domains. (Feng, Liu, Luo, & Liu, 2017) performed an analysis of over 100,000 discussion posts on Coursera. They used linear regression in combination with a Gradient Lifting Decision Tree (GBDT) to classify MOOC discussion posts. In particular, this model uses features that are independent of course content and achieved an impressive overall accuracy of 85%. (Wei, Lin, Yang, & Yu, 2017) employed a convolutional LSTM-based deep neural network to classify confusion, urgency, and sentiment in MOOC discussion forums. They learned word-level feature representations through convolutional operations, followed by learning post-level representations through an LSTM model that captures long-term temporal semantic relations. Their approach achieved an impressive 86.6% accuracy in classifying urgent and non-urgent posts.

(Guo et al., 2019) employed a hybrid deep neural network to detect urgent posts in MOOCs using pre-trained embeddings from Google News. To handle spelling errors and emoticons, they introduced a hybrid model using CNN, GRU, and Char-CNN for semantic and structural extraction. In their study, (Agrawal et al., 2015) developed a classification model to detect confusion and suggests optimal start times for video clips. The model utilized features such as bag-of-words, post metadata, and predictions for question, answer, opinion, sentiment, and urgency labels. Training was performed using standard logistic regression and for confusion label the model achieved F1-score of 77%. In other study, (Cui, & Wise, 2015) employed a binary support vector machine (SVM) to determine the relevance of question posts to course content.

Khodeir (Khodeir, 2021) utilized BERT for embedding and Bi-GRU for classification task. This model obtained 91.9% weighted F1 value. However, the proposed solution showed limited improvement, with an accuracy of 81.2% for the urgent message class, suggesting that the model is not able to effectively identify urgent messages. In a recent study, (El-Rashidy et al., 2023) introduced a four-stage model that includes coding and vectorization using pre-trained BERT, a feature aggregation method to capture data-based relationships, a CNN-based model for improved text understanding, and classification of post text using composite features. Although this approach yielded a slight improvement of 0.8% over the previous results, further improvements are needed to enhance the accuracy of the urgent class. Table 1 presents an overview of previous research results.

Table 1: Overview of previous research results.

Authors	Dataset	Approach/Model	Embedding Layer	Results/Findings
(Rashidy et al., 2023)	Stanford MOOC Posts	Feature aggregation model based on CNN	BERT	The proposed model first aggregates feature to capture data-driven relationships among token features as well as their representation. Then CNN is implemented to enhance the accuracy of text context interpretation. Finally, these combined features are utilized for post text classification. The model achieved 92.7% F1-weighted scores.
(Khodeir, 2021)	Stanford MOOC Posts	Bi-GRU	BERT	The model is created based on BERT as embedding layer and achieved 91.9% F1-weighted scores.
(Guo et al., 2019)	Stanford MOOC Posts	Hybrid CNN + GRU as well as Char-CNN	Google-news	This hybrid model was proposed to tackle with the noise of spelling errors and emoticons. The model achieved an impressed 91.8% F1-weighted scores.
(Almatrafi et al., 2018)	Stanford MOOC Posts	AdaBoost (Decision Tree)	LIWC + Term Frequency	This study tried traditional classification algorithms namely, NB, SVM, RF, AdaBoost and LR. The best 88% F1-Score was achieved by AdaBoost.
(Feng et al., 2017)	Stanford MOOC Posts	Linear Regression + gradient lifting decision tree (GBDT)	Google Word2Vec	This model classified sentiment, urgency and confusion. This model achieved 86.6% accuracy in classifying urgent posts.
(Wei et al., 2017)	More than 100,000 Coursera threads	Hybrid of CNN + LSTM		The proposed model uses features that are independent of course content and achieves an impressive overall accuracy of 85%. They also obtained that most of the posts were not related to the course contents.
(Bakharia, 2016)	Stanford MOOC Posts	NB, SVM, RF	TF-IDF with unigram features	Proposed classification model considered urgency, sentiment and confusion in different domains. This model gave good result within the domain but is difficult to generalize across different domains.
(Agrawal et al., 2015)	Stanford MOOC Posts	Logistic Regression	Bag of words with unigram features	A two-stage model is introduced: initially, the model identifies instances of confusion, followed by the application of a recommendation system to suggest brief video clips aimed at resolving that confusion. This model obtained 77% F1-scores.

3. METHOD

We present a new method to classify urgent posts in MOOC discussion forums integrating Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) layers based on BERT. CBiLSTM model aims to accurately classify MOOCs forum posts as urgent or non-urgent, addressing the challenges arising from the growing number of students and posts in MOOCs. The proposed approach follows a multi-stage process, involving data preprocessing, BERT-based data augmentation for enriching the training set with contextualized samples, BERT-based word embedding to facilitate a

deeper understanding of the text and the comprehensive description of the CBiLSTM model. The proposed approach flow is illustrated in Figure 1.

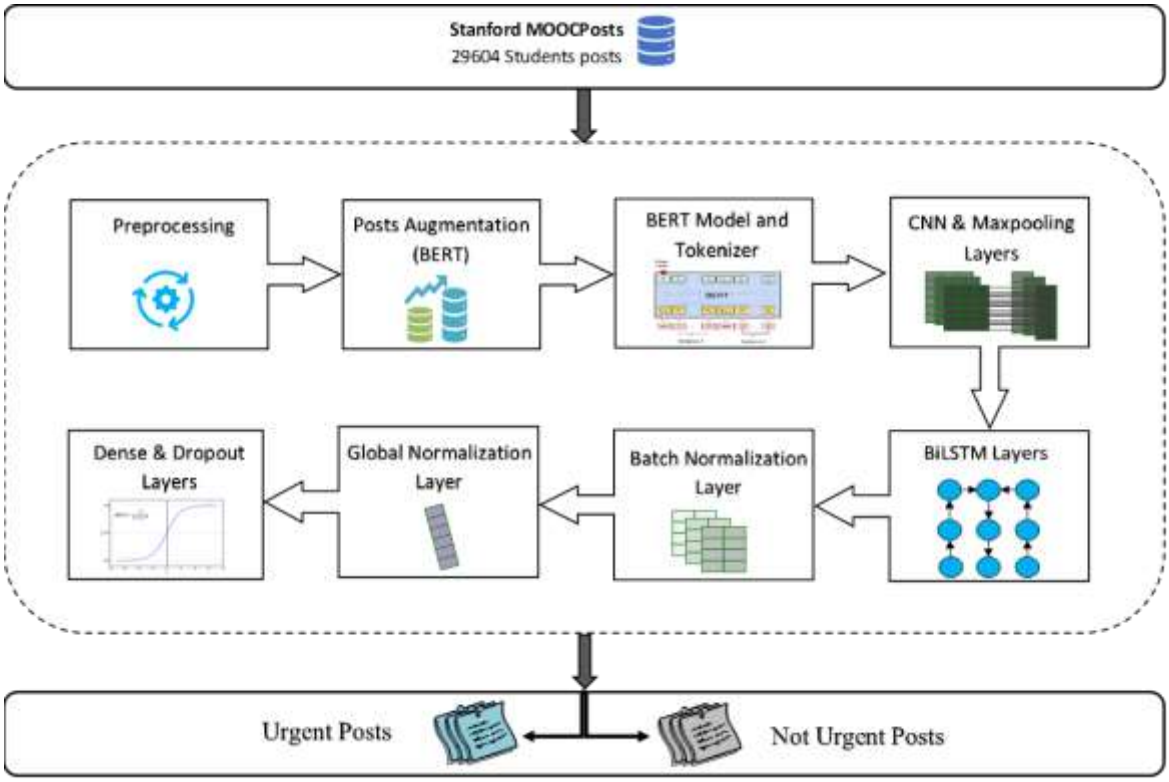

Figure 1: The proposed approach flow.

3.1 Preprocessing

In the preprocessing phase of the Stanford MOOC Posts dataset, a series of essential steps are undertaken to ensure the data consistency and quality. Firstly, all URLs (Uniform Resource Locators) were systematically removed from the text, eliminating any potential noise and irrelevant information. To standardize word forms, contractions like "won't" and "can't" are replaced with their expanded versions, such as "will not" and "can not", facilitating uniformity in text representation. Similarly, symbols like question marks and exclamation marks were substituted with a specific word, promoting a cohesive approach to text analysis.

To enhance text readability, abbreviations like "'re", "n't", "'s", and others were transformed into their corresponding full words ('are', 'not', 'is', etc.), streamlining the text for further analysis. Further, symbols such as slashes, dollar signs, and others are eliminated, ensuring that special characters do not interfere with the text's semantic meaning.

To facilitate lemmatization, the Spacy 'en_core_web_ls' model is applied, grouping inflected forms into their base or dictionary forms, leading to a more coherent and meaningful representation of words. Stop words were retained in the dataset, as their inclusion was found to improve the classification results (Agrawal, & Paepcke, 2014).

The 'course_display_name' metadata feature is integrated with student posts, displaying the course's name and domain within MOOC Posts dataset. By incorporating 'course_display_name' into the posts, a considerable improvement in the achieved results was observed (Guo et al., 2019).

3.2 Addressing Data Imbalance

Text data imbalance is a prevalent challenge in natural language processing tasks, where certain classes or categories within a dataset are significantly underrepresented compared to others. Machine learning models can be adversely affected by this imbalance, leading to biased predictions and reduced accuracy, especially for minority classes (Shaikh, Daudpota, Imran, & Kastrati, 2021). Synthetic data methods like SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) and AdaSyn (He, Bai, Garcia, & Li, 2008) effectively address statistical data imbalances. However, when applied to textual data, they face challenges with overfitting and noise. Although GANs like CycleGAN (Almahairi, Rajeshwar, Sordoni, Bachman, & Courville, 2018) have shown promise in generating synthetic numerical and image data, their suitability for textual data, including grammar, context, and semantics, requires further evaluation. We applied BERT (Bidirectional Encoder Representations from Transformers) to augment the train dataset. BERT's language understanding and contextual comprehension capabilities improved the quality of augmented data, leading to enhanced model performance. In the context of text augmentation using BERT, as seen in Figure 2, the initial step is tokenization, breaking the input sentence into discrete units like words. Subsequently, (MASK) tokens are inserted at random positions within the sentence which results in a partially masked sentence. After augmentation, the sentence is passed through the BERT model which leverages contextual information from the surrounding tokens to predict the appropriate replacements for the (MASK) tokens. Table 2 presents the original samples alongside their corresponding augmented variations.

Table 2: Original posts alongside their corresponding augmented variations.

No	Original posts	Augmented posts	Urgency
1	I love this course until the last week when try to submit work and all I get be the stupid try again come up have write in the discussion that this be happen education education one one five number how to learn math	yes I love this course until the last academic week and when try to submit work applications and instead all I get be the more stupid try again come up have write out in the joint discussion that this be happen education education one one five number how next to learn math	Urgent
2	I have the same issue take a look at the update on some platform issue discussion topic they be fix some bug education education one one five number how to learn math	I have the same issue take to a quick look then at from the news update when on some platform more issue discussion in topic they be fix some bug computer education education one one five plus number how to learn math	Not urgent

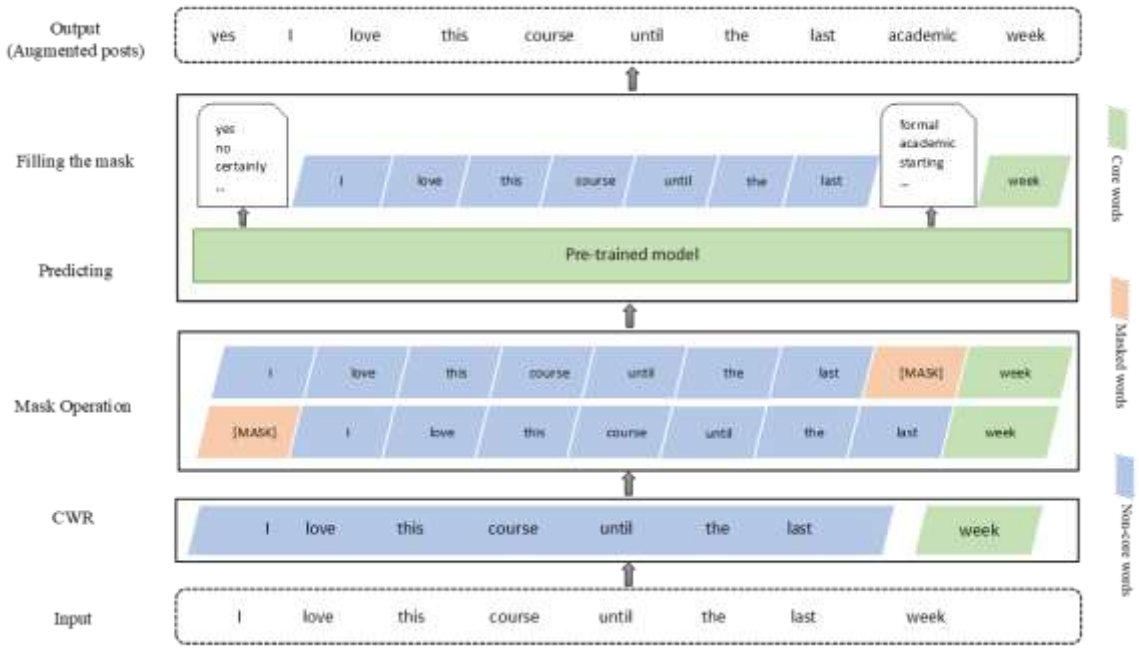

Figure 2: Internal process of BERT text augmentation.

BERT generated variations were introduced to the urgent class samples addressing the data imbalance issue, leading to notable improvements in model performance and generalization. Figure 3 illustrates the distribution of both urgent and not urgent classes following the augmentation process.

Tokenization is performed using pretrained BERT model, specifically the 'pre-trained bert-base-uncased tokenizer', which is case-insensitive and offered by the transformer's library.

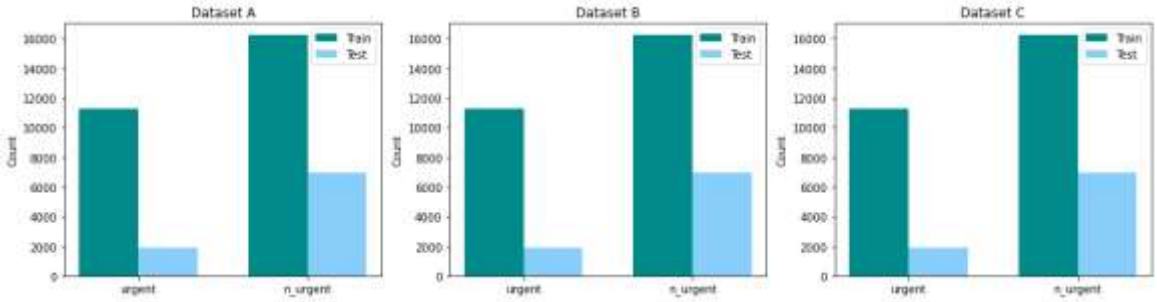

Figure 3: Distribution of urgent and not urgent classes following the augmentation process.

3.3 Embedding Layer

In NLP and neural network language models, the word embedding layer is pivotal. It is responsible for transforming words or tokens in a text into dense numerical vectors in a high-dimensional space. Before BERT, Word2Vec and Glove were widely used in NLP tasks for traditional word embeddings. BERT, as introduced by (Devlin, Chang,

& Lee, 2019), is a transformer-based model, unlike traditional approaches, BERT learns contextual embeddings. It is pre-trained on a massive corpus using next sentence prediction and masked language modeling tasks. This pre-training enables BERT to understand the semantic meaning and syntactic structure of words in various contexts. BERT-based word embeddings not only capture the semantic meaning of individual words but also their contextual significance in sentences or documents.

BERT, as shown in Figure 4, uses special tokens such as '(CLS)' and '(SEP)' to handle variable-length sequences of text. The '(CLS)' token represents the classification token and is used to achieve a fixed-size vector representation for the entire input sequence, which can be used for downstream classification tasks. The "(SEP)" token separates different sentences in the input when dealing with sentence-level tasks (Devlin et al., 2019).

In text classification, BERT encapsulates the entire sequence by utilizing the final hidden state of the special token (CLS). There are three distinct fine-tuning techniques: 1) Training the Entire Architecture: All layers of the BERT model, including task-specific layers, are updated during fine-tuning, allowing maximum flexibility and adaptation to the specific task. 2) Training Some Layers While Freezing Others: Only a subset of BERT layers is updated, while others are frozen, striking a balance between flexibility and computational efficiency. 3) Freeze the Entire Architecture: All layers, including pre-trained and task-specific, are kept frozen, making BERT function as a fixed feature extractor for subsequent tasks (Khodeir, 2021).

In this study, we load BERT pretrained model, the one we utilized for the tokenization process. The BERT model was kept frozen during the fine-tuning stage, and only the CBiLSTM component was trained to learn from BERT's representations.

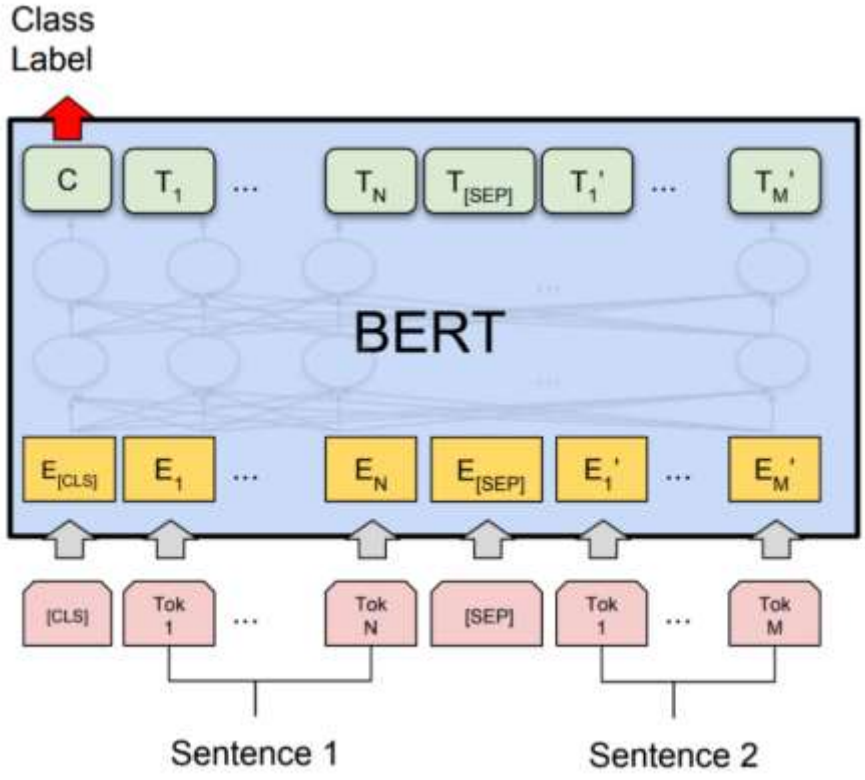

Figure 4: BERT (Language Model) (Devlin et al., 2019).

3.4 Convolutional BiLSTM (CBiLSTM) model

CNN and RNN are widely used models in sentiment analysis and text classification. Each model has distinct strengths: CNN excels at extracting local features by analyzing the spatial relationships within the text but cannot learn sequential correlations, while RNN is adept at capturing sequential correlations and extracting global features (Luan, & Lin, 2019; Du, Vong, & Chen, 2020). However, traditional RNNs suffer from issues like gradient explosion or vanishing gradient when dealing with long sequences of data. To overcome these limitations, LSTM (Du et al., 2020), an extension of RNN, was introduced. LSTM employs memory cells to handle gradient issues and capture long-term dependencies. In a traditional LSTM, data propagates through the network unidirectionally, as depicted in Figure 5-a. While this allows the LSTM to model past dependencies effectively, it may not fully leverage future context, which is important in text classification tasks.

LSTM uses forget gate (f_{gt}), input gate (i_{gt}), and output gate (o_{gt}), constructed with a sigmoid layer to control the information flow within its cells (Olah, 2020). The current cell state of the LSTM is represented as c_t. The forget gate decides what information to discard, determined by Equation 1:

$$f_t = \sigma(W_f \cdot (h_{t-1}, x_t) + b_f) \qquad 1$$

Here, w_f and b_f illustrate the forget gate's weights and bias, while x_t corresponds to the input sequence and h_{t-1} signifies the previous output state.

Next, the LSTM layer utilizes its input gate to determine which elements from the current input x_t should be incorporated into the present cell state c_t. This process involves the use of both sigmoid and tanh layers. The sigmoid layer determines updates to the current cell state, as shown in Equation 2. On the other hand, the tanh layer generates a new vector, denoted as c'_t using Equation 3, composed of the updated values. W_i and b_i represent the weights and bias of the input gate.

$$i_{gt} = \sigma(W_i.(h_{t-1}, x_t) + b_i)) \qquad 2$$
$$c't = \tanh(W_C.(h_{t-1}, x_t) + b_c) \qquad 3$$

Next, LSTM updates the previous cell state c_{t-1} with the new cell state c_t by multiplying forget gate f_{gt} values with c_{t-1}, followed by the addition of the newly computed candidate values scaled by the i_{gt}, as displayed in Equation 4. This ensures that the LSTM retains only the essential information from the current input, while discarding irrelevant or outdated information from the previous cell state.

$$c_t = f_{gt} * c_{t-1} + c'_t * i_{gt} \qquad 4$$

As result, to produce the desired output, the LSTM employs its output gate, represented by Equations 5 and 6:

$$o_{gt} = \sigma(W_o \cdot (h_{t-1}, x_t) + b_o) \qquad 5$$

$h_t = o_t \cdot \tanh(c_t)$ 6

The final hidden state h_t is subsequently passed to a fully connected layer for further processing.

Moreover, Bi-LSTM (Liu, & Guo, 2019) enhances LSTM's capabilities by incorporating two LSTM layers which simultaneously process the information bidirectionally, as depicted in Figure 5-b. The forward LSTM processes the input sequence from the first-time step to the last, while the backward LSTM processes it in reverse, from the last-time step to the first. This enables Bi-LSTM to better capture bidirectional dependencies in the data, making it particularly effective in various text classification tasks.

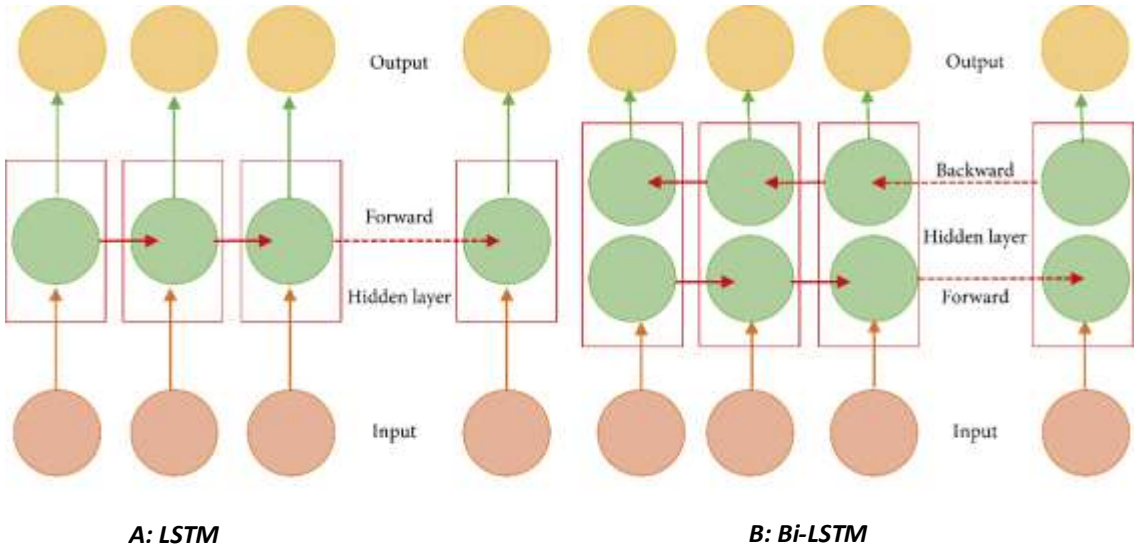

Figure 5: Depiction of LSTM model (A) alongside a Bi-LSTM model (B) (Abduljabbar, Dia, & Tsai, 2021).

We propose CBiLSTM, a hybrid deep learning model which utilizes BERT for word embedding. The model architecture comprises two main components: a CNN and BiLSTM layers. In first step, we pass the BERT-based embeddings to the CNN structure. Following the BERT embedding layer, a series of one-dimensional convolutional layers are applied. These layers employ kernel sizes of 3, 4, and 5, each with 128 filters and ReLU activation functions. These convolutions are designed to identify local patterns and features within the embedded text. After each convolutional layer, max-pooling operations with varying pool sizes (8, 6, and 3) are performed to reduce dimensionality and retain the most significant features. A dropout layer with a 0.2 rate is added after the first convolutional layer to address overfitting. Next, the feature vectors are passed through a bidirectional LSTM layer. The model incorporates three Bi-LSTM layers with hidden units of 256, 128, and 64, respectively, each set to return sequences (return sequences=True). Bi-LSTMs capture contextual information bidirectionally, enabling the model to understand the sequential dependencies within the text. To further process the output from the Bi-LSTM, another dropout layer with a rate of 0.1 is inserted to further enhance generalization. A batch normalization layer

is applied to stabilize training by normalizing activations. Following batch normalization layer, a global max-pooling layer is used which aggregates the most important features across the entire sequence, creating a fixed-length representation of the text data. The model concludes with a dense output layer featuring a single neuron using a sigmoid activation, and provides a probability value between 0 and 1. This output layer classifies the input into either the urgent or not urgent class.

For model training, the Adam optimizer is employed alongside a binary cross-entropy loss function. During model training, several callbacks are employed, including model checkpointing, early stopping, and learning rate reduction. These strategies ensure model convergence and prevent overfitting.

4. EXPERIMENT

This section provides insights into the dataset, experimental setup and evaluation metrics employed in this study.

4.1 Dataset and Experimental Setup

In this research, the experiments were conducted on the Stanford MOOC Posts dataset (Agrawal, & Paepcke, 2014), a benchmark corpus introduced by (Agrawal et al., 2015). The corpus comprises 29,604 anonymized learner forum posts from 11 Stanford University public online courses. These posts are divided into Humanities/Sciences, Medicine, and Education distinct domains, each containing 9723, 10001 and 9878 posts respectively. Each post is manually labeled across various dimensions, including assessing its urgency ranked on a scale from 1 to 7. To create a binary classification task for urgent post identification, the class labels were adjusted. Posts with urgency scores of 4 or higher were categorized as "urgent," while those with scores below 4 were labeled as "not urgent." This binary classification scheme ensures that approximately 20% of posts are classified as urgent, allowing instructors to promptly address critical cases, saving them 80% of their time, and enabling efficient management of urgent posts.

Following the approach of previous studies (Almatrafi et al., 2018; El-Rashidy et al., 2023; Khodeir, 2021), the dataset was categorized into three groups using course and domain names:

- **Group A:** The baseline scenario, where the training and test datasets remained independent across courses or domains, divided into three distinct subsets.
- **Group B:** In this case, data division was determined by the course name. Several courses were excluded from the training phase. Specifically, courses from Humanities and Medicine domains, such as Stat Learning (Winter 2014), Statistics in Medicine, and Managing Emergencies: What Every Doctor Must Know?, were selected for testing, while the Education domain had only one course, and 33% of its posts were reserved for testing in a stratified manner.
- **Group C:** In this case, domain was reserved for testing, and the classifier was trained on posts from the Medicine and Education domains. The evaluation was

performed on posts from the Humanities domain, which was held out for testing. The choice of Humanities as the domain for testing was arbitrary.

As mentioned in section 3.2, we conducted BERT-based data augmentation on the training sets of all three groups to address the dataset's class imbalance issue. This allows our model to learn more effectively and improves its performance in classifying urgent posts accurately. By achieving a balanced dataset, we aim to enhance the generalization and robustness of our model, enabling it to handle various real-world scenarios with improved precision and recall.

We employed TensorFlow and Keras libraries, utilizing Python version 3.10, to build and train our proposed model. Prior to the data being used by the model, a series of preprocessing steps were carried out to ensure data quality and consistency. To facilitate the tokenization process and word embedding, we applied the pretrained 'bert-based-uncased' model from the transformer's library. Special tokens were added to format the input sequences appropriately. As the 'bert-based-uncased' model has a constraint of a maximum input size of 512 tokens, we truncate sequences to the defined maximum length, ensuring the compatibility of our data with the BERT model. As mentioned in section 3.4, we carefully selected the optimizer, number of layers, and number of hidden units of both CNN and BiLSTM components in our model. These decisions were made after thorough experimentation and optimization to achieve the best performance.

4.2 Evaluation Metrics

In this study, we use six evaluation metrics to evaluate the model's performance: precision (Pre), recall (Rec), F1-score, F1-Weighted, Learning curve (LC) and Precision-Recall Curve (PRC). Precision measures the fraction of accurately predicted positive results by the model, while recall represents the proportion of relevant positive results correctly predicted. F1-score, a balanced metric, blends precision and recall by taking their harmonic mean to provide a single performance measure. Eq. 7, Eq. 8, and Eq. 9 present the equations for these performance metrics, illustrating their mathematical formulations.

$$Pre = TP / (TP + FP) \quad (7)$$

$$Rec = TP / (TP + TN) \quad (8)$$

$$F1\text{-}score = (2 * Pre * Rec) / (Pre + Rec) \quad (9)$$

TP (True Positives), represents correctly predicted positive cases, TN (True Negatives), denotes correctly predicted negative cases, FP (False Positives), indicates instances where the model incorrectly predicts positive cases.

Additionally, the F1-weighted score is employed to address class imbalances in the dataset by computing the weighted average of the F1-score for each class.

Learning curves are mostly employed as a diagnostic tool in machine learning, especially for incremental learning algorithms like deep learning (Anzanello, & Fogliatto, 2011). These curves provide a comprehensive evaluation of model performance by considering both the training and validation datasets. They yield two insightful curves: the Training Learning Curve and the Validation Learning Curve. The former offers a glimpse into

how effectively the model is "learning" from the training data, while the latter delves into how proficiently the model is "generalizing" its knowledge.

In the context of learning curves, it's common to employ minimization score, like loss. Smaller scores indicate more effective learning, with an ideal score of 0.0 signifying a perfect learning of the training dataset. Regularly inspecting these learning curves during training serves as a powerful tool for analyzing learning-related issues like underfitting or overfitting. In our experimental setup, we employ early stopping, a technique that involves continuous monitoring of validation set errors. Whenever an improvement is observed, we capture a snapshot of the model parameters at that point. Upon termination of the training algorithm, we retain and use these saved parameters instead of the latest ones. This approach contributes to enhance the generalization capabilities of deep neural networks.

ROC and PR curves are valuable tools for assessing probability predictions in binary classification problems (Khodeir, 2021). ROC curves are graphical representations used to illustrate the trade-off between true positive rates and false positive rates at various classification thresholds. In contrast, PR curves present the trade-off between precision and recall at different classification thresholds. When dealing with imbalanced datasets, where one class significantly outweighs the other, PR curves become more important. In such scenarios, the Precision-Recall Plot proves to be a more informative tool for assessing binary classifiers compared to the ROC plot.

To summarize model performance and facilitate comparisons between different classifiers, we turn to the Area Under the Curve (AUC) metric. AUC metric summarizes skill of the model. The precision-recall curve's baseline, denoted as $y = P/(P + N)$, where P is positives and N is negatives, acts as a reference for a no-skill classifier. Such classifier can't distinguish between classes and predicts either a random outcome or a constant class for all instances.

5. RESULT

This section features an in-depth performance analysis, comparing the proposed model with well-established baseline architectures, including standalone CNN, LSTM, and Bi-LSTM models. These baseline models provide foundational insights into the individual contributions of convolutional and sequential modeling components. It's important to note that we evaluate the performance of baseline models on augmented datasets (A, B, and C) and subsequently compared their results to our proposed model. Additionally, we compare our model performance against contemporary state-of-the-art models that have recently emerged, representing the pinnacle of advancements in urgent post classification.

5.1 Performance Comparison: CBiLSTM vs Baseline Models

In the pursuit of enhancing urgent posts classification performance, we conducted a comprehensive comparative analysis between our primary model, CBiLSTM, and foundational baseline models, namely CNN, LSTM, and BiLSTM. This evaluation encompassed a range of essential assessment tools, providing insights into the superiority of our proposed architecture.

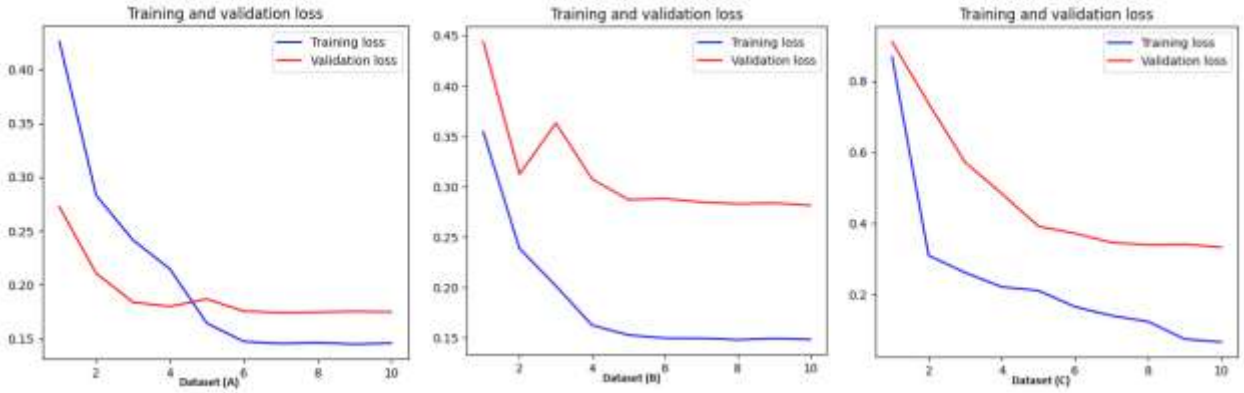

Figure 6: Learning and validation curves for CNN on datasets A, B, and C where the minimum loss values are 0.17 (Epoch 7), 0.287 (Epoch 10), and 0.331 (Epoch 10).

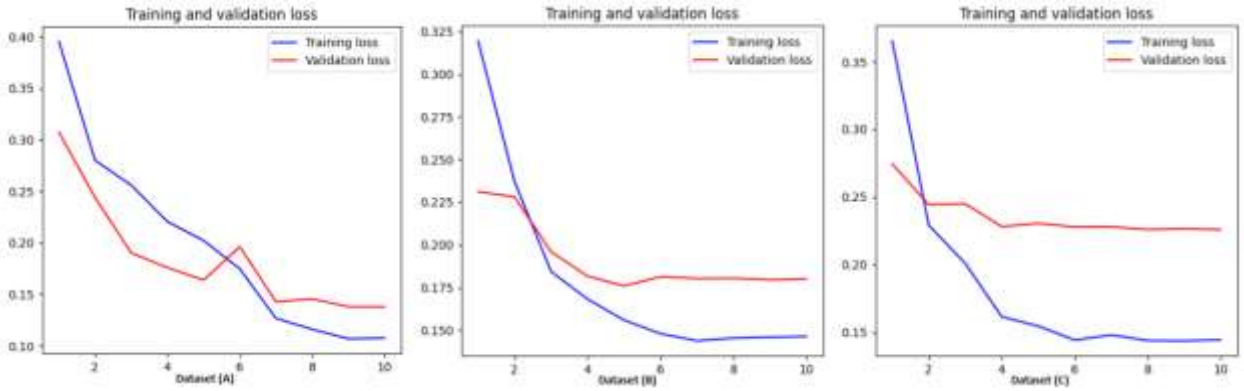

Figure 7: Learning and validation curves for LSTM on datasets A, B, and C where the minimum loss values are 0.142 (Epoch 9), 0.172 (Epoch 5), and 0.229 (Epoch 8).

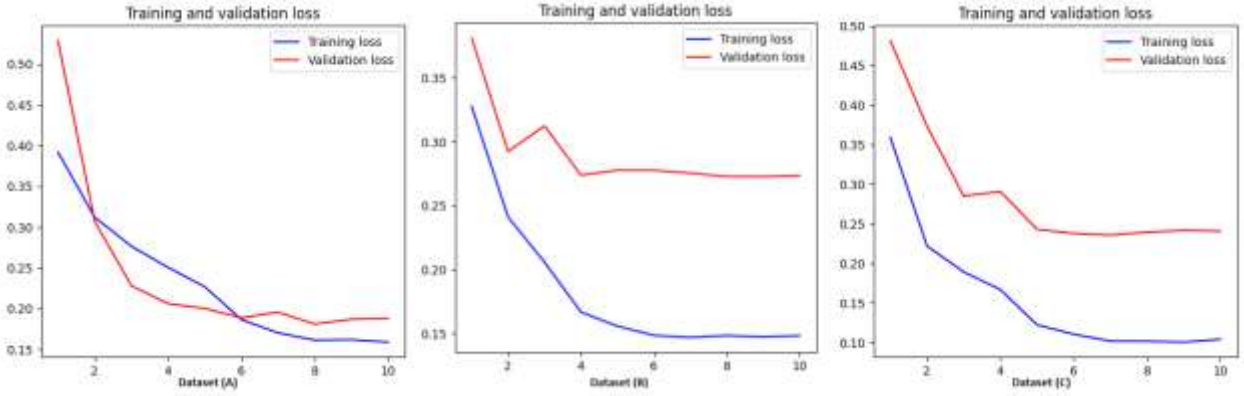

Figure 8: Learning and validation curves for BiLSTM on datasets A, B, and C where the minimum loss values are 0.181 (Epoch 8), 0.272 (Epoch 9), and 0.236 (Epoch 9).

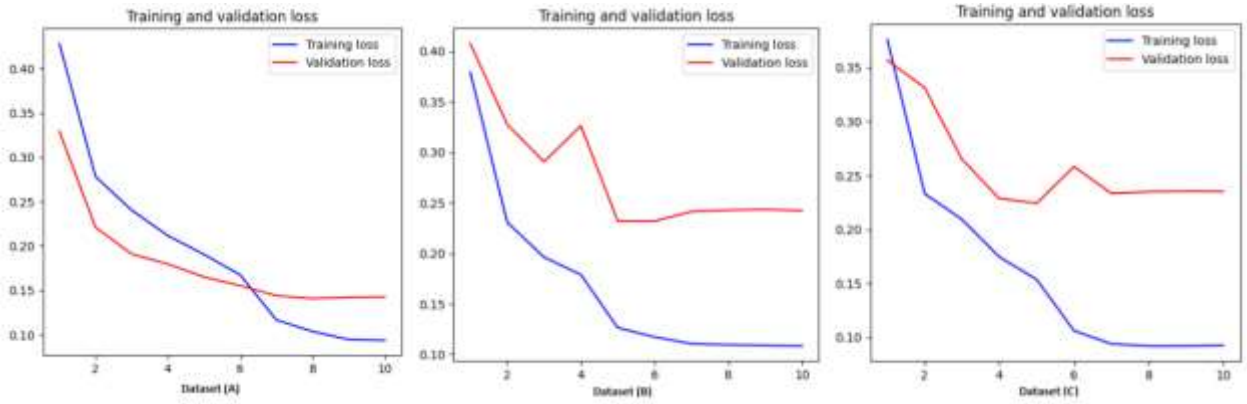

Figure 9: Learning and validation curves for CBiLSTM on datasets A, B, and C where the minimum loss values are 0.147 (Epoch 8), 0.232 (Epoch 5), and 0.226 (Epoch 5).

Learning curve analysis depicted in Figures 6 to 9 for datasets A, B, and C show how good the models are learning and generalizing. The data visualizations clearly depict that employing BERT as an embedding layer result in a substantial decrease in the training time required to meet the early stopping criteria. Additionally, the graphical representations indicate that our CBiLSTM model, when incorporating BERT as the embedding layer, achieves early stopping with minimal loss in fewer training iterations, resulting in improved efficiency and cost-effectiveness.

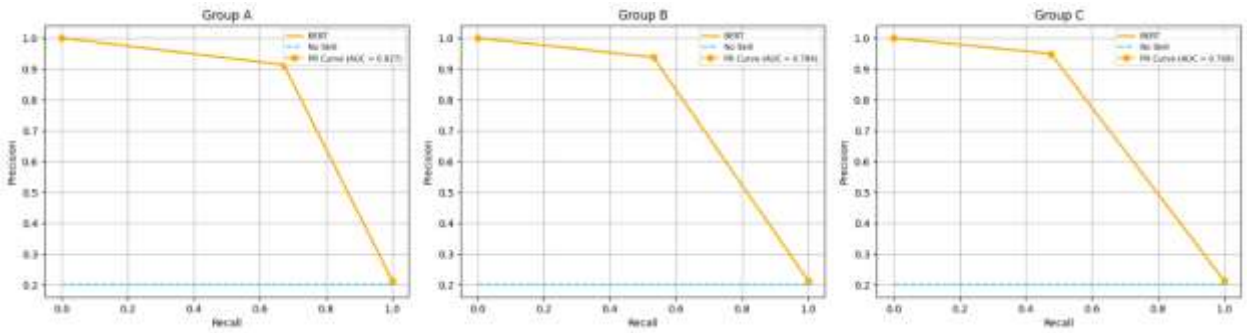

Figure 10: PR curves for CNN model using A, B, and C datasets.

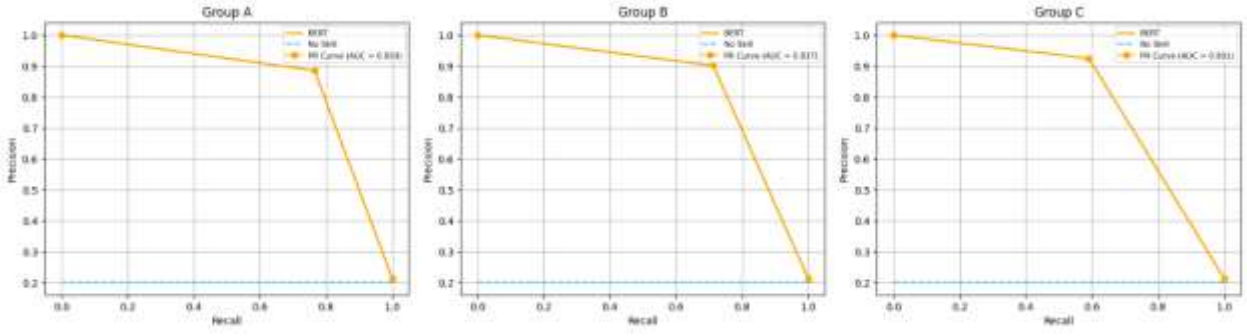

Figure 11: PR curves for LSTM model using A, B, and C datasets.

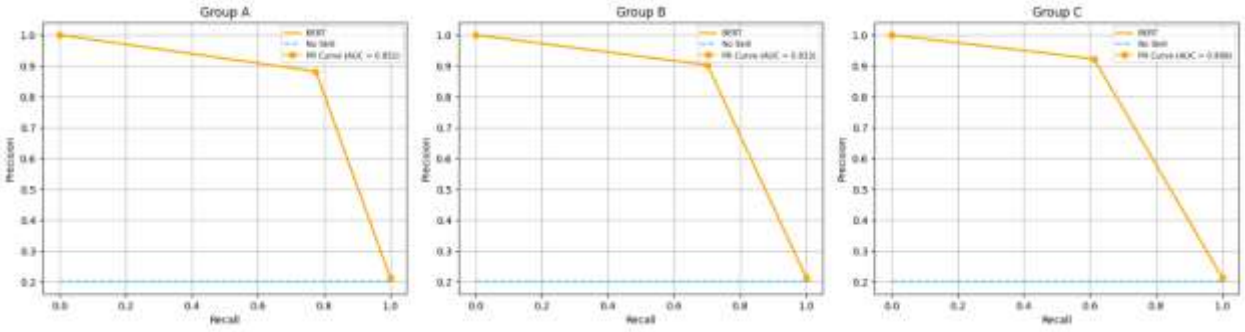

Figure 12: PR curves for BiLSTM model using A, B, and C datasets.

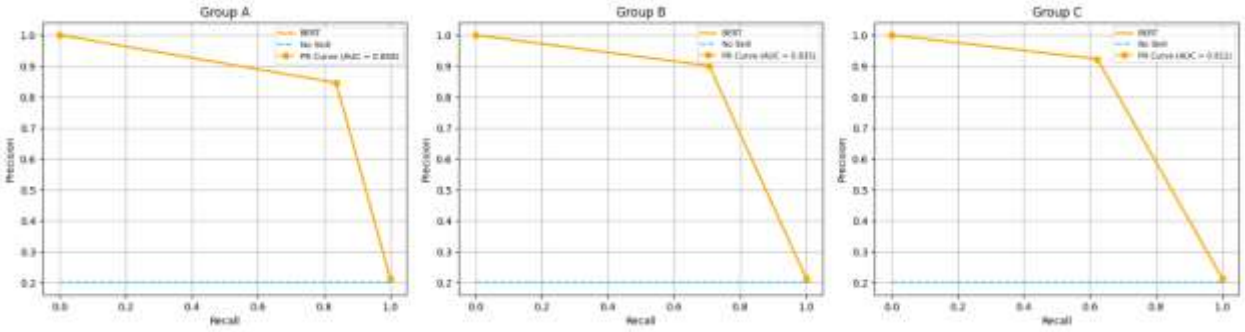

Figure 13: PR curves for CBiLSTM model using A, B, and C datasets.

The Precision-Recall curves in Figures 10 to 13 illustrate the balance between true positive rate and positive predictive value. The AUC is employed as a quantitative metric to shows the model performance. Furthermore, we employ a baseline for the precision-recall curves to show the balance between urgent and non-urgent classes. Notably, across all three datasets (A, B, and C), CBiLSTM achieves the highest AUC values, surpassing alternative techniques.

The comparative analysis of performance metrics, including Precision (ability to correctly identify positive cases), Recall (ability to capture all positive causes), F1-scores (a balanced measure of precision and recall), and F1-weighted scores, between the proposed CBiLSTM model and baseline models is presented in Tables 3, 4, and 5 for datasets A, B, and C, respectively. Notably, CBiLSTM outperforms the baseline models across A, B, and C datasets.

5.2 Performance Comparison: CBiLSTM vs State-of-the-art Models

The proposed approach's performance is compared with state-of-the-art algorithms employing three distinct A, B, and C datasets. (Almatrafi et al., 2018) presented a MOOC post classification model using TF, linguistic attributes, metadata, and AdaBoost. TF lacks consideration for word order, potentially affecting context comprehension. Linguistic features, derived from LIWC, may be impacted by misspellings and symbols, potentially influencing AdaBoost's performance. (Guo et al., 2019) proposed a method that incorporated google-news embeddings, metadata features, and a CNN-Bi-GRU

architecture. Google-news embeddings capture a word's fundamental meaning, disregarding its contextual intricacies. The CNN and Bi-GRU layers within the architecture were utilized to extract long-term dependencies within post text. (Khodeir, 2021) introduced a model that utilized preprocessing, BERT embeddings, and Bi-GRU techniques. Preprocessing encompassed the removal of stop words and special characters, which can enhance contextual comprehension but may detrimentally affect classification accuracy. The Bi-GRU component formed the classification model, emphasizing its ability to extract long-term dependencies among words effectively. Recently, (El-Rashidy et al., 2023) introduced a four-stage model that includes coding and vectorization using pre-trained BERT, a feature aggregation model to capture data-based relationships, a CNN-based model for improved text understanding, and classification of post text using composite features.

Table 3: Results from experiments on Group A.

Model	Urgent			Not Urgent			Weight F1 (%)
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
Adaboast (Almatrafi et al., 2018**)**	77	65	70	91	95	93	88
Geo et al. (Guo et al., 2019**)**	83.4	77.2	80.1	94.8	95.4	95.1	91.8
Bi-GRU (Khodeir, 2021**)**	80.8	81.5	81.2	94.9	94.7	94.8	91.9
BERT + CNN Agg (El-Rashidy et al., 2023**)**	83.6	83	83.3	95.3	95.5	95	92.7
CNN	80.7	80.4	80.5	94.7	94.8	94.8	91.8
LSTM	82.2	80.7	81.4	94.7	95.2	95	92.1
Bi-LSTM	83.7	81.3	82.5	94.9	95.6	95.2	92.5
BERT + CBiLSTM	**84.6**	**83.6**	**84.1**	**95.6**	**95.9**	**95.7**	**93.3**

Table 4: Results from experiments on Group B.

Model	Urgent			Not Urgent			Weight F1 (%)
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
Adaboast (Almatrafi et al., 2018**)**	80	65	72	90	95	92	88
Geo et al. (Guo et al., 2019**)**	80.7	79.7	80.2	93.8	**95.8**	**94.8**	**91.3**
Bi-GRU (Khodeir, 2021**)**	76	**84.7**	80.1	**95.7**	92.7	94.2	91
BERT + CNN Agg (El-Rashidy et al., 2023**)**	81.6	80.7	80.9	93.2	93.3	93.2	90
CNN	83.3	73.8	78.3	90.3	94.3	92.2	88.3
LSTM	85.7	77.8	81.5	91.9	95.1	93.5	90.2
Bi-LSTM	85.1	76.8	80.7	91.6	94.9	93.2	89.8
BERT + CBiLSTM	**86.8**	77.4	**81.8**	91.5	95.4	93.4	90.2

Table 5: Results from experiments on Group C.

Model	Urgent			Not Urgent			Weight F1 (%)
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
Adaboast (Almatrafi et al., 2018**)**	80	57	67	87	**95**	91	85
Geo et al. (Guo et al., 2019**)**	**80.7**	73.1	76.7	90.7	94.5	92.6	88.4
Bi-GRU (Khodeir, 2021**)**	76.1	83.1	79.4	94.7	92.1	93.4	90
BERT + CNN Agg (El-Rashidy et al., 2023**)**	72.4	**86.2**	78.7	**96**	91.1	93.5	90.3
CNN	74.2	83.5	78.6	94.9	91.3	93.1	89.7
LSTM	73.1	85.1	78.6	95.5	91.1	93.2	90
Bi-LSTM	73.1	85.5	78.8	95.7	91.1	93.3	90.1
BERT + CBiLSTM	76.4	85.2	**80.6**	95.5	92.2	**93.8**	**90.9**

Despite the endeavors of state-of-the-art models, the highest attained F1-weighted scores stand at 92.7%, with an 83.3% F1 score specifically for urgent class classification, as depicted in Tables 3, 4, and 5. Notably, none of these investigations have addressed the issue of dataset imbalance within the Stanford MOOC Posts dataset, a factor known to

impact classification performance and potentially introduce bias favoring larger classes (Wei, & Zou, 2019; Guo et al., 2008). Our proposed model, however, strategically addresses this imbalance through balancing the train dataset using pretrained BERT model, significantly influencing classification performance. Our proposed model has achieved remarkable results, boasting a 93.3% F1-weighted score and 84.1% F1 score for the urgent class, as demonstrated in Table 3. This represents a notable improvement of 0.6% and 0.8% over the best-performing state-of-the-art model using the group A dataset. Moreover, as shown in Table 4, on the group B dataset, our model obtained an impressive F1-urgent score boost of 1.6% compared to (Guo et al., 2019) and a 0.9% gain over the state-of-the-art model in (El-Rashidy et al., 2023). Additionally, on the group C dataset, CBiLSTM model out performs the state-of-the-art model in (El-Rashidy et al., 2023) by 0.6% in overall F1-weighted score, as well as (Guo et al., 2019) by 1.2% in the urgent post classification.

CBiLSTM model outperformed state-of-the-art algorithms using various test scenarios, obtaining higher weighted F1 scores and a balanced precision-recall for urgent posts classification (see Tables 3, 4, and 5).

6. Discussion

In this section, we discuss the key findings, implications, and contributions of our research on urgent post classification in MOOC forums.

6.1 Addressing the Urgent Post Classification Challenge

One of the central challenges in online education forums is the timely identification and response to urgent student posts. These posts can range from requests for clarification on course materials to critical issues requiring immediate instructor attention. Our research tackled this challenge by developing and evaluating the CBiLSTM model, which demonstrated remarkable performance across various evaluation metrics and outperformed both baseline models and state-of-the-art approaches. Across different dataset groupings (Group A, Group B, and Group C), our model obtained higher precision, recall, F1-scores, and F1-weighted scores comparing to baseline models. This performance suggests that our hybrid architecture effectively captures the complex patterns in forum posts. Furthermore, the learning curve analysis illustrated that CBiLSTM model, incorporating a pre-trained BERT layer as an embedding significantly reduced training time while achieving early stopping with minimal loss. This finding underscores the efficiency of our model in utilizing contextual embeddings, which is particularly important for real-time classification tasks like identifying urgent posts.

6.2 Addressing Class Imbalance

Class imbalance is a common issue in binary classification tasks and can significantly impact model performance. To overcome this challenge, we employed BERT-based data augmentation, which played an important role in improving the precision and recall of our model. By balancing the dataset, we ensured that the model's performance was not skewed toward the majority class. This approach contributes to the model's generalization and robustness, making it suitable for real-world scenarios where urgent posts are a minority.

6.3 Comparison with State-of-the-Art Model

In our comparative analysis, we observed that previous state-of-the-art models, while innovative in their approaches, did not specifically address the issue of dataset imbalance. These models achieved F1-weighted scores of up to 92.7%, with an 83.3% F1 score for urgent post classification. However, our CBiLSTM model surpassed these results with a remarkable 93.3% F1-weighted score and an 84.1% F1 score for the urgent class, as shown in Table 3. This notable improvement of 0.6% and 0.8% over the best-performing state-of-the-art model on the Group A dataset underscores the effectiveness of our approach in handling class imbalance. Additionally, on the Group B and Group C datasets, CBiLSTM consistently outperformed the state-of-the-art models by achieving higher F1-urgent scores.

7. FUTURE WORK

CBiLSTM urgent classification model, incorporating a contextual embedding layer (BERT) and BERT-based data augmentation, produced better results within the Stanford MOOC Posts datasets. However, there is potential for enhancing the model's performance. First, the contextual understanding of MOOC forum posts can be enhanced through advanced NLP techniques, potentially by incorporating transformer models like GPT-4 to provide more contextually relevant responses. Second, there is room for innovation in data augmentation, especially for scenarios with limited urgent posts, where exploring novel techniques to diversify and refine synthetic data can lead to more robust classification models. Lastly, focusing on user interface design, with an emphasis on usability studies and user feedback, is essential to create an intuitive platform that empowers both instructors and students in the MOOC environment, not only for urgent post classification but also for effective communication.

8. CONCLUSION

This paper presented a novel approach to classify urgent MOOC forum posts, addressing the challenges posed by the increasing volume of students and posts in online courses. Our model, CBiLSTM, leverages the power of BERT-based contextual embeddings and a hybrid architecture that integrates CNN with BiLSTM layers. Our method includes a multi-stage process, from data preprocessing to the integration of course metadata, and accurately tackles the issue of data imbalance through BERT-based data augmentation. Balancing the dataset through BERT-based augmentation played a critical role in enhancing the model's performance. This strategic step effectively addressed the challenges posed by class imbalance, leading to more accurate and robust classification results. We have shown the effectiveness of our model by comparing it to baseline models, such as standalone CNN, LSTM, and BiLSTM, as well as state-of-the-art approaches in the domain.

Across various datasets representing different scenarios, our CBiLSTM model consistently outperforms baseline models and demonstrates remarkable improvements over existing state-of-the-art algorithms. Notably, our model obtained higher F1-weighted scores and balanced precision-recall for urgent post classification.

Acknowledgement
This work was supported by Shahid Rajaee Teacher Training University.

Credit author statement
Mujtaba Sultani: Conceptualization, Data curation, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft.
Negin Daneshpour: Conceptualization, Methodology, Writing – review & editing.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest
None: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Availability of data and material
Data are available on justified request to the corresponding author.

REFERENCES

Anzanello, M. J., & Fogliatto, F. S. (2011) ‘‘Learning curve models and applications: Literature review and research directions. *International Journal of Industrial Ergonomics*, *41*(5), 573–583.

Agrawal, A., & Paepcke, A. (2014). The stanford MOOC Posts dataset. *Accessed: Dec*, *15*, 2020.

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). YouEDU: Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips. *International Educational Data Mining Society*.

Almatrafi, O., Johri, A., & Rangwala, H. (2018). Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. *Computers & Education, 118,* 1-9.

Almahairi, A., Rajeshwar, S., Sordoni, A., Bachman, P., & Courville, A. (2018, July). Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International conference on machine learning* (pp. 195-204). PMLR.

Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.

Abduljabbar, R. L., Dia, H., & Tsai, P. W. (2021). Unidirectional and bidirectional LSTM models for short-term traffic prediction. *Journal of Advanced Transportation*, *2021*, 1-16.

A Decade of MOOCs: A Review of Stats and Trends for Large Scale Online Courses in 2021. EdSurge.’’ https://www.edsurge.com/news/2021-12-28-a-decade-of-moocs-a-review-of-stats-and-trends-for-large-scale-online-courses-in-2021.

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM computing surveys (CSUR)*, *49*(2), 1-50.

Bakharia, A. (2016, April). Towards cross-domain MOOC forum post classification. In *Proceedings of the third (2016) ACM conference on learning@ scale* (pp. 253-256).

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Cui, Y., & Wise, A. F., (2015, March). Identifying content-related threads in MOOC discussion forums. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 299-303).

Chang, C. Y., Lee, S. J., Wu, C. H., Liu, C. F., & Liu, C. K. (2021). Using word semantic concepts for plagiarism detection in text documents. *Information Retrieval Journal*, *24*, 298-321.

Devlin, J., Chang, M. W., & Lee, K. (2019, June). Google, KT, Language, AI: BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171-4186). (Online). Available: http://arxiv.org/abs/1810.04805

Du, J., Vong, C. M., & Chen, C. P. (2020). Novel efficient RNN and LSTM-like architectures: Recurrent and gated broad learning systems and their applications for text classification. *IEEE transactions on cybernetics*, *51*(3), 1586-1597.

El-Rashidy, M. A., Farouk, A., El-Fishawy, N. A., Aslan, H. K., & Khodeir, N. A. (2023). New weighted BERT features and multi-CNN models to enhance the performance of MOOC posts classification. *Neural Computing and Applications*, 1-15.

Feng, Y., Chen, D., Zhao, Z., Chen, H., & Xi, P. (2015, August). The impact of students and TAs' participation on students' Academic performance in MOOC,'' in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (pp. 1149-1154).

Feng, L., Liu, G., Luo, S., & Liu, S. (2017). A transferable framework: Classification and visualization of mooc discussion threads. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24* (pp. 377-384). Springer International Publishing.

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008, October). On the class imbalance problem. In *2008 Fourth international conference on natural computation* (Vol. 4, pp. 192-201). IEEE.

Guo, Z. X., Sun, X., Wang, S. X., Gao, Y., & Feng, J. (2019). Attention-based character-word hybrid neural networks with semantic and structural information for identifying of urgent posts in MOOC discussion forums. *IEEE access*, *7*, 120522-120532.

He, H., Bai, Y., Garcia, E. A., & Li, S., (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). Ieee.

Hone, K. S., & El Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study. *Computers & Education, 98*, 157–168.

Kumar, V., Choudhary, V., & Cho, E. (2020). Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Khodeir, N. A. (2021). Bi-GRU urgent classification for MOOC discussion forums based on BERT. *IEEE Access*, *9*, 58243-58255.

Liyanagunawardena, T. R., Adams, A. A., & Williams, S. A. (2013). MOOCs: A systematic study of the published literature 2008-2012. *International Review of Research in Open and Distributed Learning*, *14*(3), 202-227.

Luan, Y., & Lin, S. (2019, March). Research on text classification based on CNN and LSTM. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)* (pp. 352-355). IEEE.

Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, *337*, 325-338.

Olah, C. (2020). Understanding lstm networks, August 2015. *URL https://colah. github. io/posts/2015-08-Understanding-LSTMs. Accessed on*, *10*. (Online). http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, *5*(4), 1-29.

Shah, D. (2016). Monetization over massiveness: Breaking down MOOCs by the numbers in 2016. EdSurge.

Sun, X., Guo, S., Gao, Y., Zhang, J., Xiao, X., & Feng, J. (2019, March). Identification of urgent posts in MOOC discussion forums using an improved RCNN. In *2019 IEEE world conference on engineering education (EDUNINE)* (pp. 1-5). IEEE.

Shaikh, S., Daudpota, S. M., Imran, A. S., & Kastrati, Z. (2021). Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. *Applied Sciences*, *11*(2), 869.

Wei, X., Lin, H., Yang, L., & Yu, Y. (2017). A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information*, *8*(3), 92.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Xue, J., & Chen, Y. (2022, July). The principle and implementation of sentiment analysis system. In *International Conference on Artificial Intelligence and Security* (pp. 28-39). Cham: Springer International Publishing.

Zhenghao, C., Alcorn, B., Christensen, G., Eriksson, N., Koller, D., & Emanuel, E. J. (2015). Who's benefiting from MOOCs, and why. *Harvard Business Review*, *25*(1), 2-8.

Zhang, C., Chen, H., & Phang, C. W. (2018). Role of instructors' forum interactions with students in promoting MOOC continuance. *Journal of Global Information Management (JGIM)*. 26(3), 105-120.