Extracting Urgent Questions from MOOC Discussions: A BERT-based Multi-Output Classification Approach

ABSTRACT

Online discussion forums are widely used by students to ask and answer questions related to their learning topics. However, not all questions posted by students receive timely and appropriate feedback from instructors, which can affect the quality and effectiveness of the online learning experience. Therefore, it is important to automatically identify and prioritize student questions from online discussion forums, so that instructors can provide better support and guidance to the students. In this paper, we propose a novel hybrid Convolutional Neural Network (CNN) + Bidirectional Gated Recurrent Unit (BiGRU) multi-output classification model, which can perform this task with high accuracy and efficiency. Our model consists of two outputs: the first one classifies whether the post is a question or not, and the second one classifies whether the classified question is urgent or not urgent. Our model leverages the advantages of both CNN and BiGRU layers to capture both local and global features of the input data, as well as the Bidirectional Encoder Representations from Transformers (BERT) model to provide rich and contextualized word embeddings. The model achieves an F1-weighted score of 94.8% when classifying whether the posts are questions or not, and obtains an 88.5% F1-weighted score while classifying the question into urgent and non-urgent. Distinguishing and classifying urgent student questions with high accuracy and coverage can help instructors provide timely and appropriate feedback, a key factor in reducing dropout rates and improving completion rates.

Keywords

Discussion Forums, MOOC, Urgent Question, BERT, CNN, BiGRU

1. INTRODUCTION

In the realm of education, the value of student feedback has long been recognized as a vital tool for assessing the quality of instructional processes and improving overall learning experiences. Traditional educational institutions often rely on periodic surveys to gather student perspectives, seeking insights into the effectiveness of course delivery, instructor performance, and the achievement of learning objectives [1]. These surveys, typically conducted at mid-term or end-term intervals, provide quantitative data that can be analyzed statistically and can contribute to ongoing improvements in educational practices.

Over the years, the rise of online learning, particularly through Massive Open Online Courses (MOOCs), has brought a transformative shift in the dynamics of student feedback. The scalability and accessibility inherent in MOOCs attract a diverse global audience which results in a need for innovative approaches to feedback collection and analysis [2]. While the traditional model of surveys remains applicable, the sheer volume of participants in MOOCs, often with high student-to-teacher ratios, demands more efficient and real-time feedback mechanisms. Within the MOOC ecosystem, discussion forums serve as dynamic spaces where learners engage in meaningful interactions, share experiences, and seek assistance [3]. These forums, embedded in the MOOC infrastructure, play a pivotal role in shaping the learning environment. However, the decentralized nature of these interactions presents a challenge in harnessing the wealth of unstructured data generated by student posts, limiting the ability to extract meaningful insights [4]. While these forums offer a rich source of information, a significant challenge involves efficiently extracting and classifying urgent student questions, a crucial aspect that has been overlooked in prior research.

The need for effective MOOC feedback analysis becomes paramount in the face of reported attrition rates and the challenges posed by questionnaire biases [5]. Real-time monitoring and comprehension of student feedback are crucial for reducing disengagement and providing instructors and course designers with actionable insights for course improvement [6]. Prior research has explored the analysis of MOOC feedback, revealing shortcomings in traditional methods like surveys and closed-ended questions, which struggle to capture the depth of student sentiments [7]. However, the potential lies in the unstructured data within discussion forums, where students freely express their thoughts, opinions, and, significantly, urgent questions that demand immediate attention.

In the existing literature, the specific task of urgent question extraction within MOOC feedback analysis remains notably absent. While previous studies have classified posts into urgent and non-urgent categories [8 - 13], none have focused specifically on identifying urgent student questions. It's noteworthy to emphasize that within the broader category of urgent posts, there is a substantial amount that includes not only immediate inquiries but also valuable student answers, opinions, and suggestions [14]. Recognizing this, the efficient extraction of urgent questions becomes even more critical, ensuring prompt responses to address learners' immediate concerns [10]. The need for urgent question extraction is paramount in MOOCs due to the huge number of participants and limited

instructor resources. Timely identification and response to urgent questions are crucial for student engagement, satisfaction, and ultimately, course completion rates. Moreover, previous studies often face limitations in capturing the contextual meaning of words and addressing imbalanced datasets, hindering their effectiveness in distinguishing urgent posts that demand immediate attention.

Recently, the dropout rate has become a significant concern in the MOOC context [15]. Previous studies highlighted the correlation between active forum participation, successful question-answering, and reduced dropout rates [16 - 18]. This underscores the importance of extracting urgent questions and providing timely answers to student queries which can boost retention and engagement, and importantly help decrease the dropout rates. We propose a BERT-based hybrid multi-output deep learning model named CBiGRU, explicitly designed to extract and classify urgent student questions within MOOC discussion forums. The proposed model is a combination of Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU) layers. In the CBiGRU model, the CNN layers effectively handle the high dimensionality of the input texts, and Bi-GRU layers explore feature context bidirectionally. CBiGRU not only distinguishes questions from non-questions but also provides a nuanced classification based on urgency, ensuring timely and targeted support for learners. The key contributions of our paper include:

1. Based on our literature review and analysis, this is the first research paper to extract and classify urgent student questions in the education domain, particularly within the context of MOOCs.
2. We propose CBiGRU, a BERT-based multi-output hybrid deep learning model to extract urgent student questions from MOOC discussion forums.
3. We implement a BERT-based data augmentation technique to address dataset imbalance and enhance the model's ability to generalize across diverse educational contexts.
4. We employ the BERT pre-trained model as an embedding layer to capture the contextual meaning of words within the MOOC discussion forums.
5. The paper helps instructors, course designers, and policymakers to enhance decisions on course participation, and improvement, with a focus on urgent student concerns.

The following sections present the structure progression of our paper. We provide a brief review of the relevant literature in Section 2. Then, we describe the proposed method we used in Section 3. Section 4 covers the data source, experimental design, and evaluation criteria we applied in this study. We present and discuss our findings in Section 5, and conclude with some remarks in Section 6.

2. RELATED WORK

In the realm of MOOCs and their discussion forums, substantial research has been conducted to extract valuable insights from user-generated content. The focus of these studies ranges from student sentiment analysis to urgent post classification and to the more nuanced aspect of opinion and suggestion mining. While existing literature has made significant strides in understanding and categorizing user contributions, a notable gap exists in the specific context of urgent question extraction within MOOC discussion forums.

Several studies have focused on urgent post-classification within MOOCs, each contributing unique insights. Almatrafi and his co-authors [10] addressed this challenge by utilizing metadata, linguistic features, and traditional machine learning algorithms, with AdaBoost producing a better result. [8] proposed a model to classify posts by urgency, sentiment, and confusion in different domains. However, this model is not very general and works well only for specific courses or domains. [19] conducted an extensive analysis of discussion posts on Coursera. They used linear regression and Gradient Lifting Decision Tree (GBDT) to classify MOOC discussion posts. The proposed model is independent of course content and achieved an overall 85% accuracy. Wei and his colleagues [9] employed a convolutional LSTM-based deep neural network to classify confusion, urgency, and sentiment in MOOC discussion forums. They learned word-level features with convolution operations and captured long-term semantic relations in posts through LSTM layers. The proposed model obtained 86.6% accuracy in classifying urgent and non-urgent posts and showed the potential of deep learning models.

Guo et al. [11] introduced a hybrid deep neural network using pre-trained embeddings from Google News to detect urgent posts in MOOCs. The model handled spelling errors and emoticons and emphasized semantic and structural extraction. Agrawal and his co-authors [20] proposed a classification model to identify confusion and recommend optimal start times for video clips. The paper leveraged features like bag-of-words, and post metadata, and made predictions across various labels. Furthermore, [12] utilized BERT for embedding and Bi-GRU as a classification algorithm and obtained a 91.9% weighted F1 score. The paper demonstrated the effectiveness of advanced embedding techniques but highlighted the need for further improvement in accurately identifying urgent posts. El-Rashidy and his co-authors [13] presented a four-stage model that incorporated pre-trained BERT model as an embedding layer, a feature aggregation method, a CNN-based model, and classification of urgent posts using composite features. The model enhanced text understanding and classification accuracy.

In the realm of opinion and suggestion mining, Almatrafi & Johri [14] analyzed MOOC discussion forum posts to summarize participants' opinions and identify suggestions for improvement. The study utilized sentiment analysis and rule-based techniques. This study marked a significant step forward in understanding participant opinions and extracting aspect-based suggestions, particularly in the educational context. Macina and her co-authors [21] suggested routing questions to willing and knowledgeable participants. They noted that certain MOOC questions necessitate instructor responses. Cui and Wise [22]

employed a binary support vector machine (SVM) to determine the relevance of question posts to course content. The paper obtained that only a small proportion (28%) of the learners' questions were content-related.

Despite these notable contributions, the literature reveals a substantial gap concerning the extraction and classification of urgent questions within MOOC discussion forums. Urgent questions are a unique subset of forum interactions and require immediate responses to address learners' immediate concerns. This paper introduces a BERT-based CBiGRU multi-output hybrid deep learning model. Unlike previous studies that primarily focused on general urgent post classification, our model is explicitly designed to extract and classify urgent student questions within MOOC discussion forums. Table 1 provides a summary of prior research findings.

Table 1:Summary of prior research findings

Authors	Dataset	Embedding Layer	Approach/Model	Evaluation Metrics	Results/Findings
[13]	Stanford MOOC Posts	BERT	Feature aggregation model based on CNN	Precision, Recall, F1 score, F1 weighted score, PR-curves, AUC	The proposed model achieved 92.7% F1-weighted scores for urgent post classification.
[14]	Stanford MOOC Posts		Rule-based	Precision, Recall, F1 score, Kappa	The model achieved 31% F1-score and overall, 0.41 Kappa scores for suggestion mining and sentiment analysis.
[12]	Stanford MOOC Posts	BERT	Bi-GRU	Precision, Recall, F1 score, F1 weighted score, PR-curves, AUC	The model obtained 91.9% F1-weighted scores for urgent post classification.
[11]	Stanford MOOC Posts	Glove (Google-news)	Hybrid CNN + GRU as well as Char-CNN	Precision, Recall, F1 score	The model achieved 91.8% F1-weighted scores.
[10]	Stanford MOOC Posts	LIWC + Term Frequency	NB, SVM, RF, LR and AdaBoost (Decision Tree)	Precision, Recall, F1 score, F1 weighted score, Kappa	The best 88% F1-Score was achieved by AdaBoost.
[19]	Stanford MOOC Posts	Google Word2Vec	LR + gradient lifting decision tree (GBDT)	Accuracy, AUC	The model achieved 86.6% accuracy in classifying urgent posts.
[9]	More than 100,000 Coursera threads		Hybrid of CNN + LSTM	Accuracy	The proposed approach achieved overall 85% accuracy.
[8]	Stanford MOOC Posts	TF-IDF with unigram features	NB, SVM, RF	Accuracy	The model gave good result within the domain but is difficult to generalize across different domains.
[20]	Stanford MOOC Posts	Bag of words with unigram features	Logistic Regression	Precision, Recall, F1 score, Kappa	The model obtained 77% F1-scores.

[22]	27739 Coursera threads	Bag or words (Unigram & Bigrams)	SVM	Precision, Recall, Accuracy, Kapp	The paper obtained that only a small proportion (28%) of the learners' questions were content-related.

3. METHOD AND PROPOSED APPROACH

In this section, we detail the methodology employed for the extraction and classification of urgent student questions within MOOC discussion forums. We propose a CBiGRU multi-output hybrid deep learning model that applies a two-step classification process. Initially, the model identifies whether a student's post is a question or not, and then, in the second step, it further classifies questions into urgent or non-urgent categories. The subsections cover data preprocessing, data balancing, and the development of the proposed multi-output classification approach. Figure 1 represents the sequential flow of these operations, providing a concise overview of the proposed approach.

3.1 Preprocessing

In the realm of applying Deep Learning to text, several preprocessing steps are necessary to ensure meaningful analysis. As shown in Figure 2, this paper employs the following steps:

1. **Symbolic Element Standardization:** We substitute symbolic elements such as question marks, exclamation marks, and ampersands, with specific words that maintain semantic integrity.
2. **URLs Removal:** URLs are removed to eliminate any potential noise and irrelevant information.
3. **Contractions Standardization:** Contractions such as "won't" and "can't" are replaced with "will not" and "can not" for uniform text representation.
4. **Special Character Removal:** Special characters such as slashes and dollar signs are removed to avoid interference with semantic meaning.
5. **Lemmatization with Spacy Model:** We apply the Spacy 'en_core_web_ls' model for lemmatization which groups inflected forms into their base or dictionary forms.
6. **Stop Words Retention:** We retain stop words in the dataset, aligning with previous findings [23], to enhance the classification results.
7. **Metadata Feature Integration:** We integrate the 'course_display_name' metadata feature with student posts, following practices from related research [11], to significantly improve overall results.

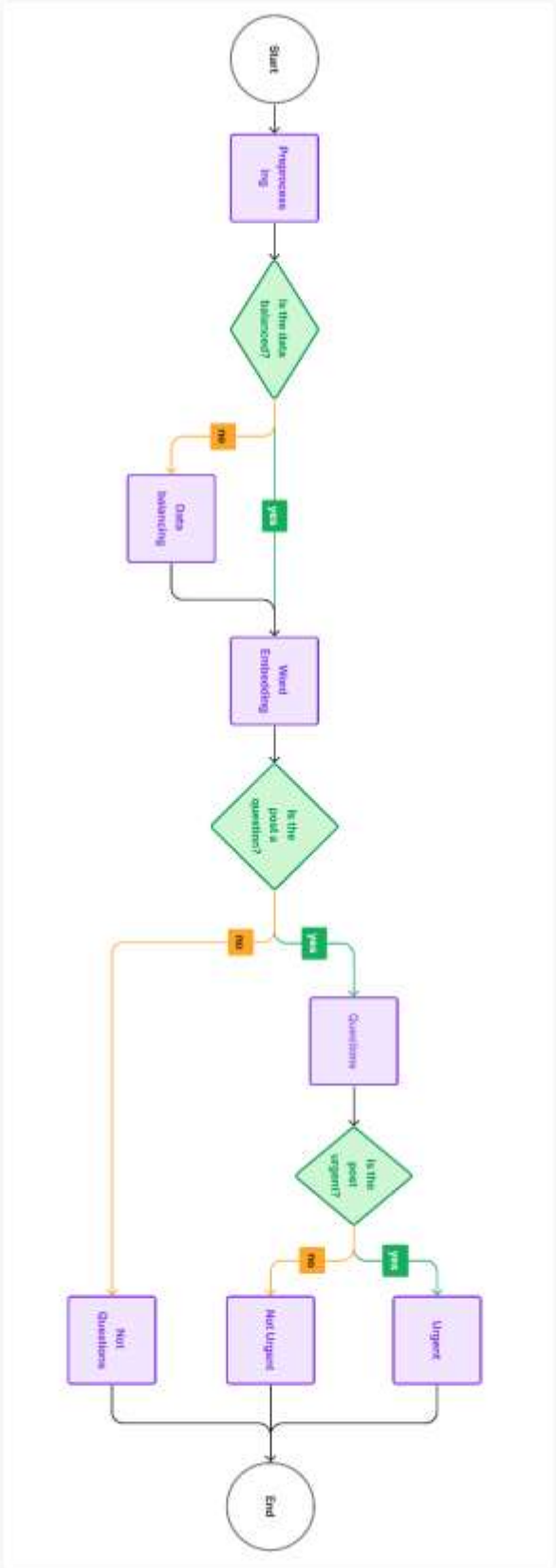

Figure 1: The proposed approach sequential flow

3.2 Dataset Balancing

In the realm of machine learning, a dataset is considered imbalanced when certain classes or categories within it are significantly underrepresented compared to others. This scenario is particularly common in natural language processing tasks, where specific outcomes might occur less frequently. The imbalance introduces a skewed distribution, with some classes having an excess of instances while others are notably scarce [24 - 27]. The imbalance in a dataset can have profound implications for machine learning models. When one class dominates the dataset, models trained on such imbalanced data might exhibit biased behavior, favoring the majority class and leading to suboptimal performance in minority classes. This is a critical concern as it can result in reduced accuracy, sensitivity, and overall model effectiveness, especially for the classes that are already underrepresented [24].

Table 2: Original posts alongside augmented variations

No	Original Posts	Augmented Posts	Question	Urgent
1	i would like to start the year off by share some of these video with my student especially about mindset and make mistake i think it be important for they to be aware of their habit and try to shift their way of think education education one one five number how to learn math	i would like to start the year off by share together some of up these video with my student especially about mindset and make mistake i think let it be important and for how they to be socially aware of their habit and their try to shift from their way of being think education education one one five number how to naturally learn math	No	No
2	Should i take comfort in score one three correct question i try to do the algebra with ould a visual and make hash of two three education education one one five number how to learn math	Should i must take comfort in score one others three correct in question i try to now do the simple algebra with ould a visual and make hash of numbers two numbers three elementary education education one one one five number three how to learn math	Yes	No
3	Youtube be block from my country if i can not not get access to the lecture i be afraid i have to stop now if you be the instructor or ta can you give i some clarification of the video download issue thank you humanity science state learn winter two zero one four	Youtube be block from my entire country if i either can not not get access to the lecture i be afraid i might have to stop now look if you be the art instructor or ta own can you give i got some clarification of the video download issue thank you to humanity the science state learn winter two zero one four	Yes	Yes

To mitigate the challenges associated with data imbalance, we employ a BERT-based data augmentation strategy. BERT (Bidirectional Encoder Representations from Transformers), is known for its language understanding and contextual comprehension capabilities. The augmentation process begins with tokenization, breaking input sentences into discrete units. [MASK] tokens are then randomly inserted within the sentence. The sentences with these partially masked tokens are processed through the

BERT model, leveraging contextual information from surrounding tokens to predict suitable replacements for the [MASK] tokens. Table 2 shows the original samples alongside their augmented variations and illustrates the effectiveness of BERT in generating diverse and contextually relevant data. While we acknowledge the challenges associated with data imbalance, we emphasize the importance of maintaining a nuanced balance between different labels. To balance the 'question' class, we carefully monitor the impact on the balance between 'urgent' and 'not-urgent' labels, recognizing the intrinsic connection between these aspects.

Figure 3 visually depicts the distribution of classes post-augmentation, showing a more balanced representation. In Figure 2, the 'A: Question' axis represents a binary scale, where 0 denotes that the posts are categorized as not-questions, and 1 indicates that the posts are questions. Meanwhile, on the 'B: Urgency' axis, the Likert scale ranges from 1 to 7, reflecting the urgency level of the posts. A rating of 1 suggests that the post is not urgent, while a higher score, such as 7, signifies a greater sense of urgency in the content.

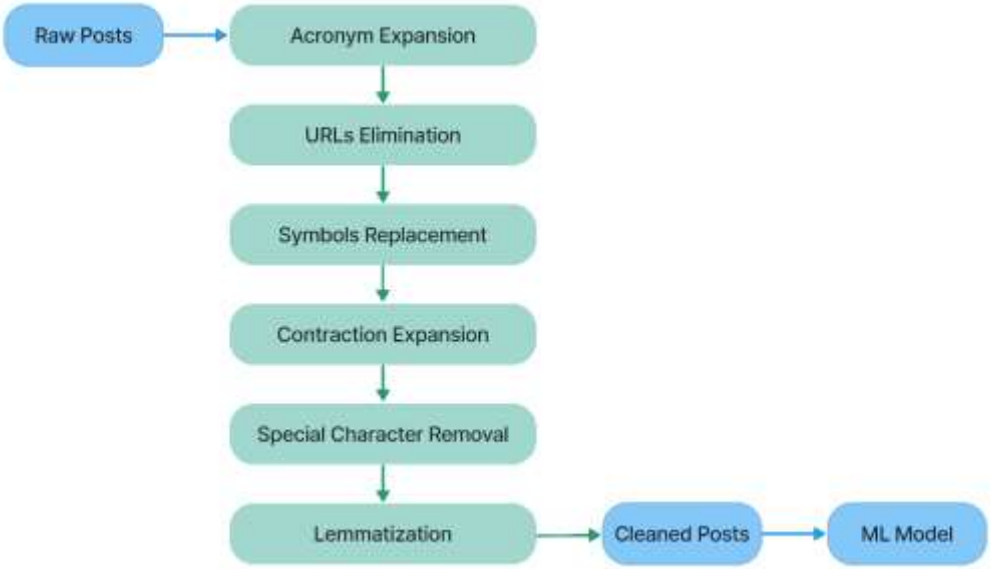

Figure 2: Preprocessing steps

3.3 Proposed Multi-output Classification Approach

In this section, we delve into word embedding and alongside that, we define the proposed CBiGRU multi-output hybrid deep learning model.

3.3.1 Word Embedding

Word embedding methods are fundamental for capturing semantic and syntactic intricacies in language. The vector space model, an early approach, represents words based on their syntactic aspects within a document or across a corpus. This method employs a fixed-dimensional vector, and the one-hot representation sets

the corresponding entry as 1 if the word is present in the text [12]. To refine this representation, Term Frequency (TF) considers the frequency of words, while Inverse-Document Frequency (IDF) penalizes common words and rewards rare ones, and results in the TFIDF score that captures both syntactic and semantic information [28].

Word embeddings, grounded in the distributional hypothesis, construct fixed-length, dense, and distributed representations, acknowledging the significance of a word's relationship with its context. These embeddings play a pivotal role in natural language processing tasks, enhancing the understanding of the interplay between words in diverse contexts [29, 30]. Another approach to word embedding is through prediction-based models, such as the Neural probabilistic model, Word2Vec, and FastText. The neural probabilistic model learns embeddings by predicting a word's context through one-hot encoded vectors and back-propagation. Word2Vec is implemented as Continuous Bag-Of-Words (CBOW) or Skip Gram (SG) [31, 32]. In CBOW, the model predicts surrounding context words using the current word, while in continuous SG, the roles are reversed, with the current word predicting each word in its context. FastText is a Facebook-provided toolkit [33], and enhances the Skip Gram model. FastText captures sub-word structure, and diverse word senses, and provides improved representations for rare or unseen words. Count-based models, like GloVe, leverage global word-context co-occurrence counts to derive word embeddings. GloVe, in particular, uses neural methods to create expressive and dense word vectors by considering global statistics in the language modeling task [12].

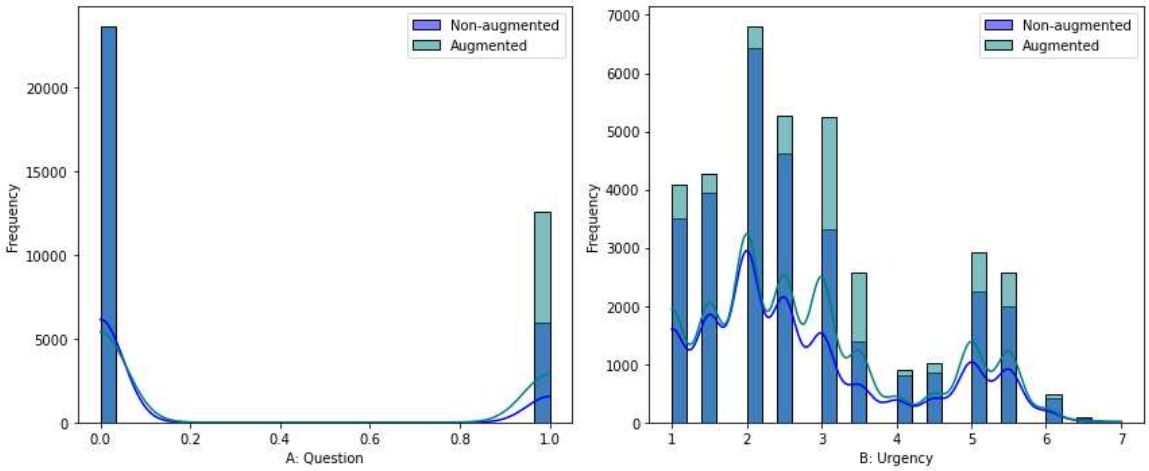

Figure 3: Comparative Distribution of Questions and Urgency Labels Pre and Post Augmentation

In addition to prediction-based models, the BERT introduces context-aware word representations and allows words to have different embeddings based on their context. In BERT, as defined by [34], each word's embedding is determined by its context within the sentence, allowing for nuanced and context-dependent representations. BERT introduces special tokens such as [CLS] and [SEP] to

enhance its functionality. The [CLS] token represents the entire sentence and is employed for classification tasks, while [SEP] marks the separation between sentences in input sequences. BERT employs three fine-tuning strategies: 1) Updating the complete architecture, 2) selectively updating specific layers, and 3) keeping BERT as a frozen feature extractor [12]. In our study, we utilize the 'bert-based-uncased' pre-trained BERT model for tokenization and word embedding. During fine-tuning, the BERT model remains frozen, and only the Bi-LSTM component is trained to learn from BERT's representations.

3.3.2 CBiGRU multi-output Deep Learning Model

In the realm of sentiment analysis and text classification, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) play an important role. CNNs excel in local feature extraction and spatial relationships but cannot learn sequential correlations [35], whereas RNNs are adept at capturing sequential correlations and extracting global features [36]. Traditional RNNs, while proficient in capturing sequential correlations, suffer from vanishing gradients when dealing with lengthy data sequences. LSTM was introduced as an extension of RNN to address this limitation [36]. It employs memory cells, forget gates, input gates, and output gates to control information flow and capture long-term dependencies effectively. LSTM's unidirectional data flow, however, may not fully leverage future context, which is crucial in tasks like text classification. Bi-LSTM [37] enhances LSTM by incorporating two parallel layers that process data bi-directionally. One layer processes the input sequence from the first step to the last, while the other processes it in reverse, from the last step to the first.

Gated Recurrent Units (GRU), proposed as an alternative to LSTM, maintains gating units for information flow modulation without separate memory cells [38]. GRU can better perform compared to LSTM, being easier to train and significantly improving training efficiency. GRU incorporates two gated units: the reset gate (r_t) controls what information to discard, and the update gate (z_t) determines which elements from the current input should be integrated into the current cell state. The GRU model uses the sigmoid function to convert data into values in the range of 0 to 1, serving as the gated signal. The computation of the reset and update gates is defined in Eq. 1 and Eq. 2:

$$r_t = \sigma\,(w_r \cdot [h_{t-1}, x_t]) \quad (1)$$

$$z_t = \sigma\,(w_z \cdot [h_{t-1}, x_t]) \quad (2)$$

Where, x_t and h_{t-1} represent the input and the previous hidden state, respectively. The weights w_r and w_z are associated with the reset and update gates, respectively. The logistic sigmoid function σ converts the computed values into the range of 0 to 1.

The subsequent steps involve utilizing the reset gate to obtain reset data (h'_{t-1}) and perform memory updates, as shown in Eq. 3 and Eq 4. The reset data, combined

with the input, is scaled using a hyperbolic tangent (tanh) activation function to scale data to the range of -1 to 1.

$$h_t = \tanh\left(w_{h'} \cdot [h'_{t-1}, x_t]\right) \quad (3)$$

$$h'_{t-1} = h_{t-1} \cdot r_t \quad (4)$$

The memory updating stage combines the update gated state (h'_t) with the previous hidden state (h_{t-1}), considering the selective forgetting and remembering mechanisms using Eq. 5:

$$h_t = z_t \cdot h'_t + (1 - z_t) \cdot h_{t-1} \quad (5)$$

The updated gating signal (z_t) in the range of 0 to 1 determines the balance between remembering and forgetting. If z_t is closer to 1, more data is remembered, while if closer to 0, more information is forgotten. This dual role of the gate enables GRU to selectively forget and remember, contributing to efficient information processing.

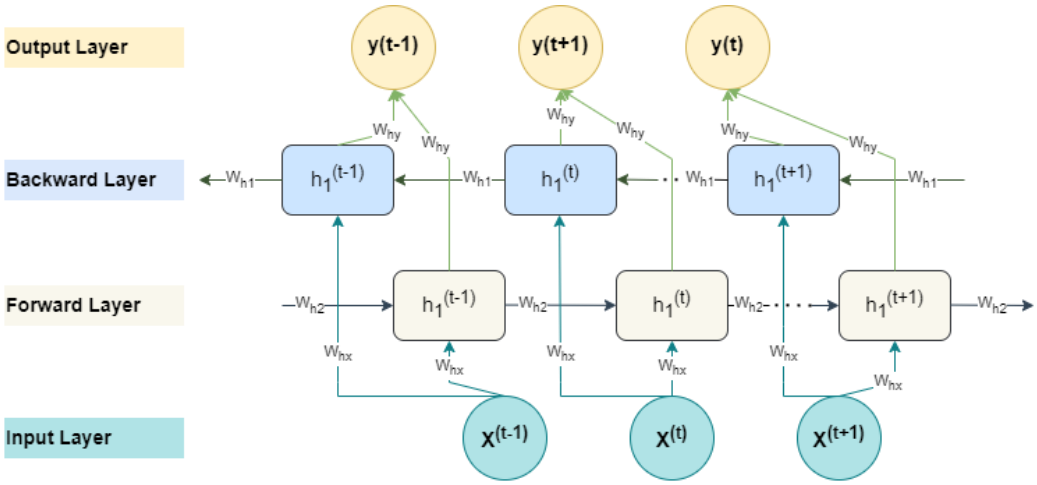

Figure 4: Bi-GRU architecture

Traditional GRU models are unidirectional, processing data only from the forward sequence. To enhance the model's ability to capture context information bidirectionally, Bi-GRU was introduced. As shown in Figure 4, Bi-GRU combines forward GRU with reverse GRU, utilizing two independent hidden layers to process data simultaneously from both forward and reverse sequences. This bidirectional approach enables the model to learn more context information and ultimately improves classification accuracy [39].

The model architecture consists of multiple layers to extract and classify urgent questions. It begins with BERT-based embeddings derived from input token IDs and their masks. To capture local features, these embeddings then pass through two one-dimensional convolutional layers each with 128 filters and kernel sizes of 3 and 6 respectively. Following each CNN layer, max-pooling operations with varying pool sizes 8 and 3 are performed to reduce dimensionality (Figure 5). To prevent overfitting, a dropout layer with a rate of 0.2 is added after the first convolutional layer. Furthermore, the CNN output is passed into two Bi-GRU layers, each consisting of 128 units and returning sequences. Dropout regularization is applied to mitigate overfitting. The output from the second Bi-GRU layer is used for question classification, passing through a flattened layer and a dense layer with a sigmoid activation function. To focus on samples classified as questions, the output from the dropout layer is filtered using element-wise multiplication with the question classification output. Subsequently, filtered questions fed into two additional Bi-GRU layers for urgency classification, followed by dropout regularization and flattening. The final urgency classification output is obtained using a dense layer with a sigmoid activation function. The entire multi-output model is compiled with the Adam optimizer and binary cross-entropy loss functions for both question and urgency outputs. Training is performed with a batch size of 200, and the BERT layers are kept non-trainable, ensuring that the pre-trained embeddings remain fixed and do not get updated during the training process. The model is trained for 10 epochs, with validation data and callbacks for model checkpointing, early stopping, and learning rate reduction. Figure 6 shows the model architecture.

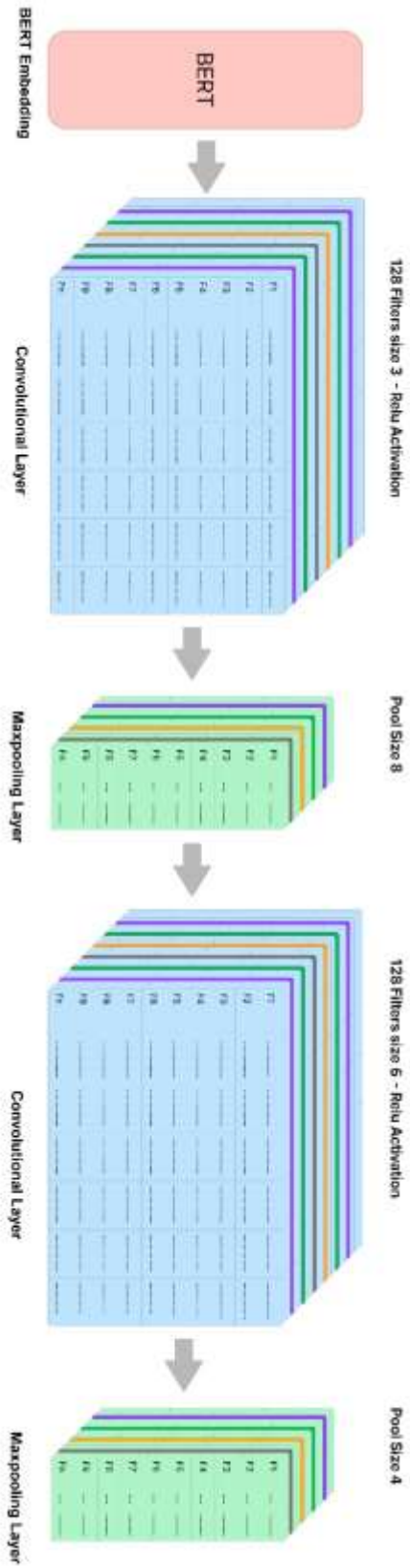

Figure 5: Capturing spatial features from input data using multiple filters

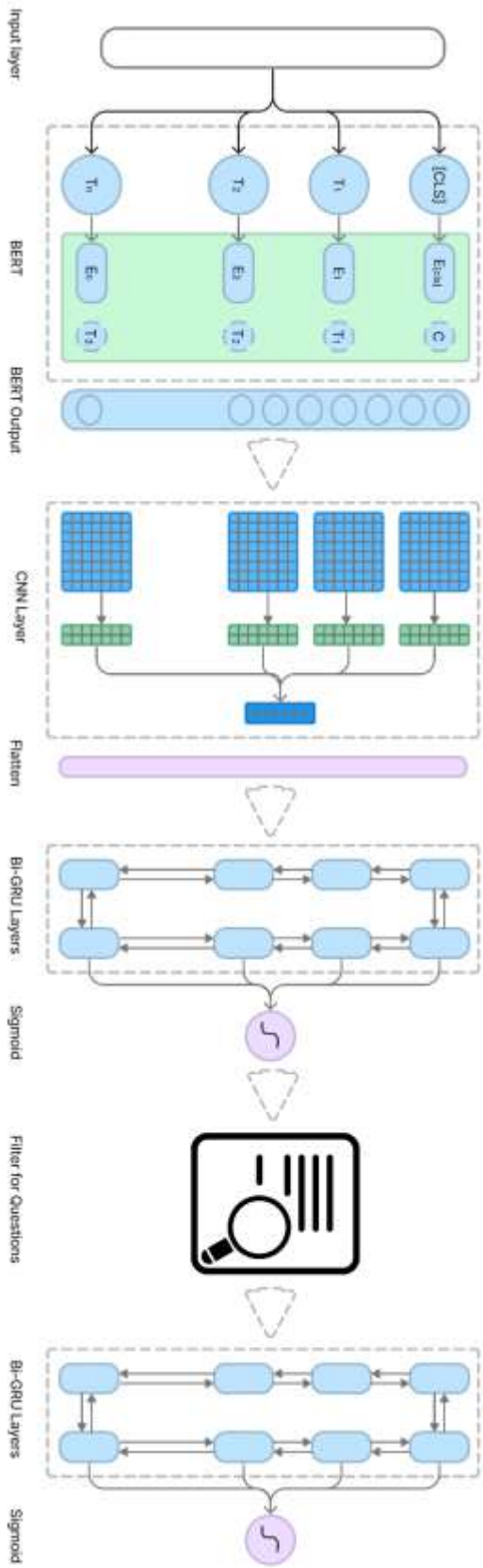

Figure 6: Proposed model architecture

4. EXPERIMENT

In this section, we provide an overview of the dataset, the experimental setup, and the evaluation metrics used in this study.

4.1 Dataset

In this paper, the experiments are conducted using the Stanford MOOC Posts dataset, a standardized benchmark corpus introduced by [20]. The dataset contains 29,604 anonymized forum posts from 11 public online courses from Stanford University, stratified across three domains Humanities/Sciences, Medicine, and Education. As shown in Figure 7, manual labeling was applied across multiple dimensions, including the categorization of posts into student questions and non-questions, coupled with an evaluation of urgency on a scale ranging from 1 to 7.

To perform a binary classification task for the identification of student urgent questions, the urgency class labels are adjusted based on urgency scores. Posts that have scores of 4 or higher are labeled as 'urgent,' while those scoring below 4 are labeled as 'not urgent.' This binary classification paradigm ensures that around 15% of posts are earmarked as urgent questions. This methodological refinement serves instructors to promptly address critical cases, streamline their workload, and facilitate effective management of urgent questions.

4.2 Experimental Setup

In our experimental setup, the model is built and trained using TensorFlow and Keras libraries in Python version 3.10. We performed several preprocessing steps on the data before feeding it to the model to ensure data quality and consistency. We used the pre-trained ‘bert-based-uncased’ model from the transformer’s library for tokenization and word embedding. Special tokens were added to format the input sequences properly. To make sure our data is compatible with the BERT model, we limit the sequences to a maximum length of 512 tokens, as this is the maximum input size of the ‘bert-based-uncased’ model. After extensive experimentation, we selected the optimizer, the number of layers, and the number of hidden units for Bi-GRU layers in the proposed multi-output hybrid classification model.

4.3 Evaluation Metrics

To evaluate the model's performance, we employ Precision (PR), Recall (RC), F1-score, weighted F1-score, Learning Curve (LC), and Precision-Recall Curve (PRC) analysis tool [12]. Despite leveraging BERT-based data augmentation to enhance dataset balance, some imbalance persists. Therefore, in addition to PR, RC, and F1-score, weighted F1-score, LC, and PRC are suitable evaluation metrics to analyze the model's performance. PR is the ratio of true positives (TP) to the total number of predicted positives, measuring

the model's accuracy in identifying the positive class. RC is the ratio of TP to the total number of actual positives, measuring how completely the model captures the positive class. F1-score is the harmonic mean of PR and RC. It balances both metrics and gives a single score that reflects the model's performance on the positive class. The mathematical formulations of these performance metrics are shown in Eq. 6, Eq. 7, and Eq. 8.

$$PR = TP / (TP + FP) \quad (6)$$

$$RC = TP / (TP + TN) \quad (7)$$

$$F1\text{-score} = (2 * Pre * Rec) / (Pre + Rec) \quad (8)$$

TP is the number of correct positive predictions, TN (True Negatives) is the number of correct negative predictions, and FP (False Positives) is the number of incorrect positive predictions. The weighted F1 score is the weighted average of the F1 scores for each class. It takes into account the class imbalance and gives more weight to the classes with fewer samples.

LC offers a visual representation of the model's learning process over varying dataset sizes. It helps us to understand how the model learns from the data and whether it suffers from underfitting or overfitting [40]. PRC shows the trade-off between precision and recall for different threshold values. It provides nuanced insights into the model's performance, especially in imbalanced scenarios [12]. To summarize the model's effectiveness and compare classifiers, we use the Area Under the Curve (AUC) metric. AUC summarizes how well the model performs overall. The PRC's baseline, denoted as y=P/(P+N), sets a standard for a no-skill classifier that cannot distinguish between different classes.

	Text	Opinion(1/0)	Question(1/0)	Answer(1/0)	Sentiment(1-7)	Confusion(1-7)	Urgency(1-7)	CourseType	forum_post_id	course_display_name
0	Interesting! How often we say those things to others without really understanding what we are saying. That must have been a powerful experience! Excellent!	1	0	0	6.7	2.0	1.5	Education	52251772c5618a0000915	EducationEDUC115NHow to Learn Math
1	"What is Algebra as a Math Gamer" or are you just saying you create games that incorporate algebra."	0	1	0	4.0	5.0	3.5	Education	52074be9935dfc0e000005e	EducationEDUC115NHow to Learn Math
2	I like the idea of my Mds principal who says 'Smart doesn't mean easy, smart means working hard" and incorporating the idea of making mistakes into working hard."	1	0	0	5.5	3.0	2.5	Education	52052d2d81fee0a9000071	EducationEDUC115NHow to Learn Math

Figure 7: A sample of Stanford MOOC Posts dataset

5. RESULT AND DISCUSSION

The proposed multi-output classification approach presented in this paper, designed to extract and classify urgent student questions, marks a novel contribution to the field. Therefore, a direct comparison of our proposed model with previous studies may not be conceptually sound due to the different goals of this paper. However, concerning the binary classification of student comments in MOOC discussion forums, our proposed model achieved better results compared to previous studies. The best F1-weighted score reported by previous studies for the binary urgency classification task is 92.7%, as shown in Table 3. In comparison, our multi-output model excels with a 94.5% F1-weighted score for the initial classification determining whether a post is a student question (Table 4). Furthermore, in the subsequent task of classifying the urgency level of student questions, our model achieved an F1-weighted score of 88%, as shown in Table 5.

Table 3: Results of the previous studies for urgency classification

Model	Urgent			Not Urgent			Weight F1 (%)
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
Adaboast [10]	77	65	70	91	95	93	88
Geo et al. [11]	83.4	77.2	80.1	94.8	95.4	**95.1**	91.8
Bi-GRU [12]	80.8	81.5	81.2	94.9	94.7	94.8	91.9
BERT + CNN Agg [13]	83.6	83	**83.3**	95.3	95.5	95	**92.7**

We compared our hybrid CBiGRU model with six baseline models: CNN, LSTM, BiLSTM, a hybrid of CNN and BiLSTM, GRU, and BiGRU. As mentioned, the learning curve can help us to understand how well the model is learning from the data and whether it is overfitting or underfitting. Figures 8 to 14 show the learning curves of the proposed model and the baseline models for both classification tasks. As we can see, CBiGRU has the lowest training and validation loss among all the models for both outputs, indicating that it can fit the data well and generalize to new data. CBiGRU model also converges faster than the other baseline models, reaching the minimum loss at epoch **8**, while the other models have taken **8** or more epochs. This shows that the proposed model is more efficient and robust than the baseline models for both outputs.

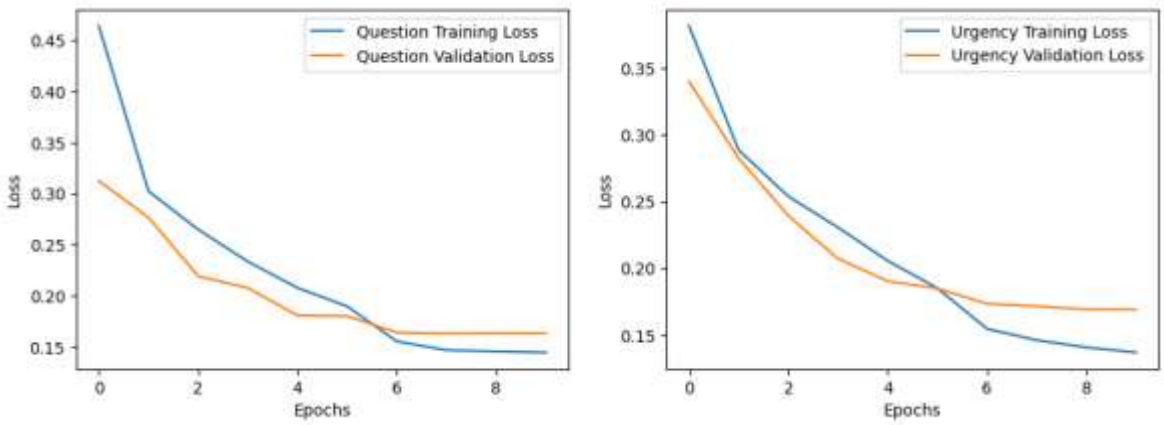

Figure 8: Learning & validation curves with minimum validation loss and corresponding Epochs for CNN model: Question (loss 0.163, Epoch 8), Urgency (loss 0.169, Epoch 9)

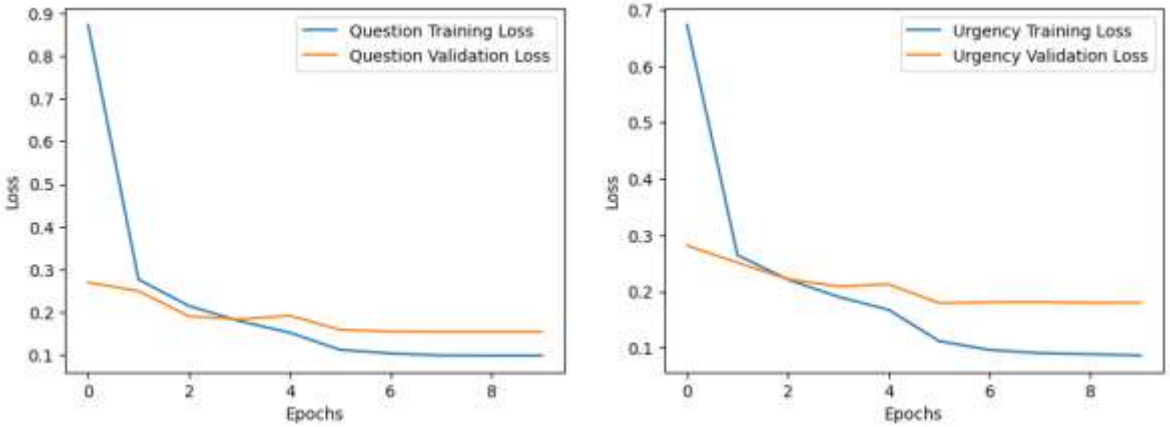

Figure 9: Learning & validation curves with minimum validation loss and corresponding Epochs for LSTM model: Question (loss 0.153, Epoch 8), Urgency (loss 0.179, Epoch 5)

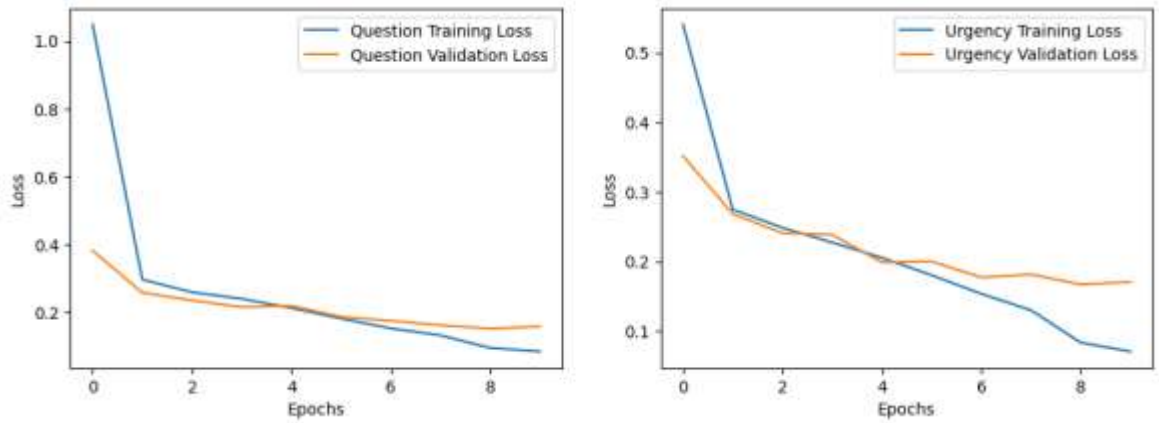

Figure 10: Learning & validation curves with minimum validation loss and corresponding Epochs for GRU model: Question (loss 0.151, Epoch 8), Urgency (loss 0.167, Epoch 8)

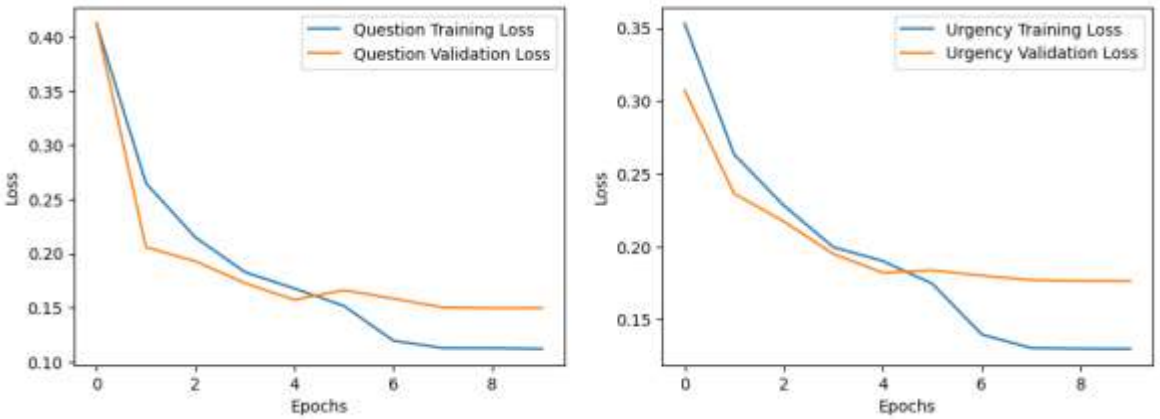

Figure 11: Learning & validation curves with minimum validation loss and corresponding Epochs for Bi-LSTM model: Question (loss 0.150, Epoch 7), Urgency (loss 0.176, Epoch 8)

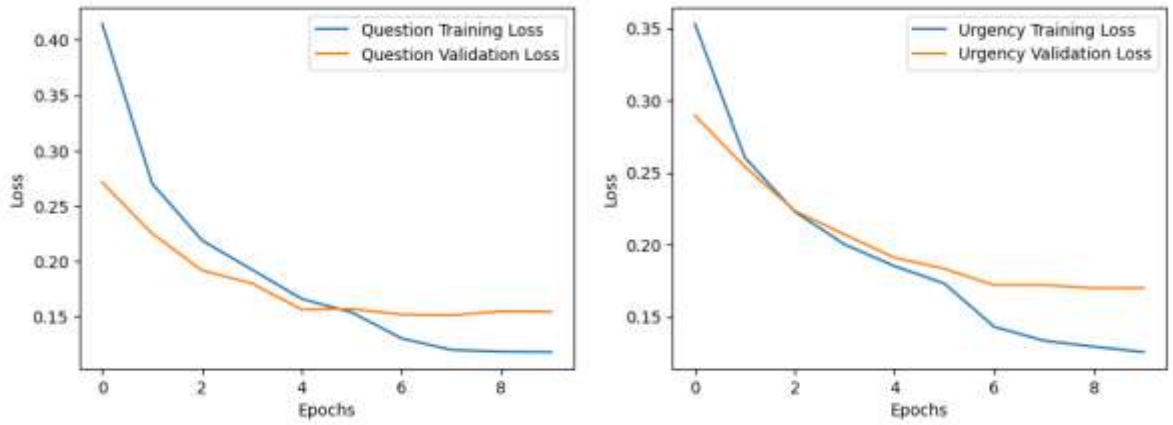

Figure 12: Learning & validation curves with minimum validation loss and corresponding Epochs for Bi-GRU model: Question (loss 0.151, Epoch 7), Urgency (loss 0.170, Epoch 8)

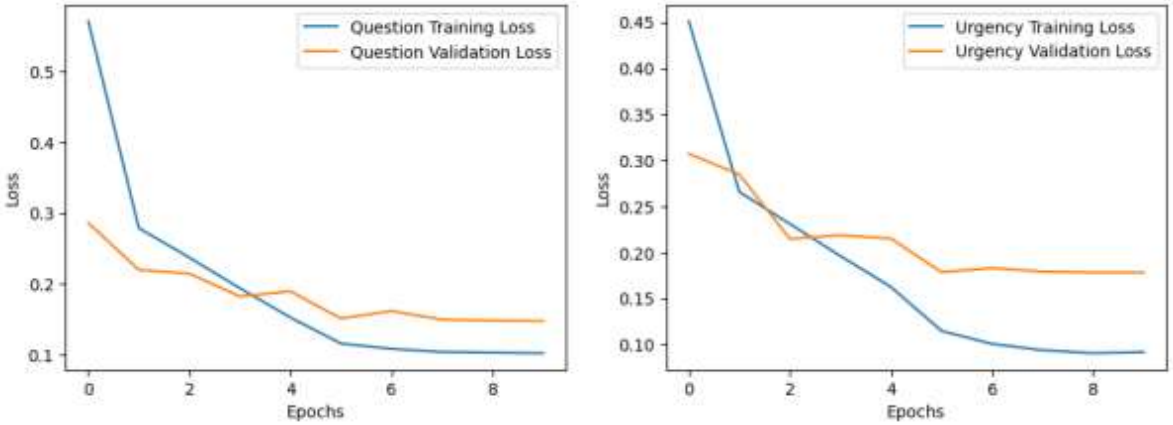

Figure 13: Learning & validation curves with minimum validation loss and corresponding Epochs for CNN + Bi-LSTM model: Question (loss 0.151, Epoch 7), Urgency (loss 0.171, Epoch 7)

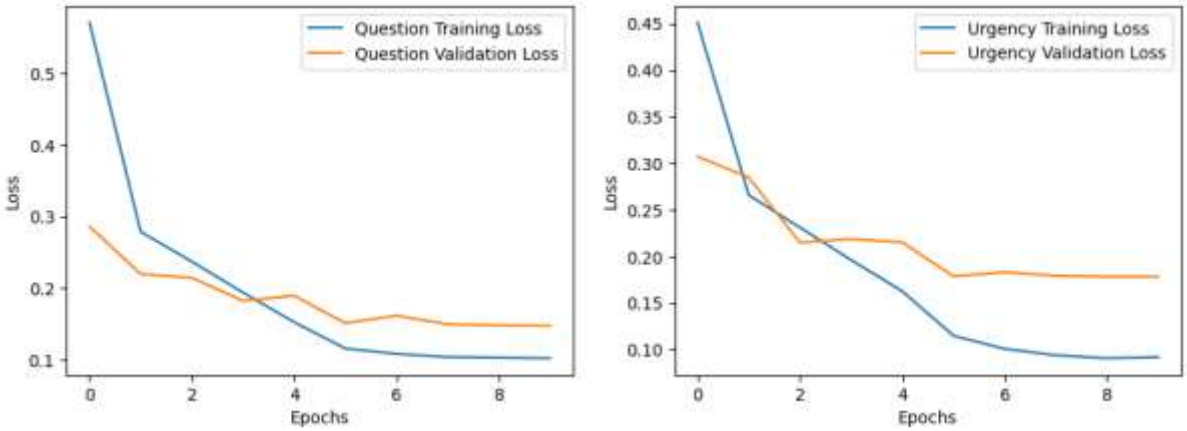

Figure 14: Learning & validation curves with minimum validation loss and corresponding Epochs for CNN + Bi-GRU model named CBiGRU: Question (loss 0.149, Epoch 5), Urgency (loss 0.161, Epoch 5)

Table 4 shows the precision, recall, F1-score, and F1-weighted score of the CBiGRU model and the baseline models on the test set for the first output, which is the question or not question classification. As we can see, the CBiGRU model achieved better performance among all baseline models which indicates that it can correctly identify the posts that are questions with high confidence and coverage. The proposed model outperformed the second-best model (CNN + Bi-LSTM) by **0.6%** in the F1 score for the question class and **0.5%** in the F1-weighted score for the entire classification task. Table 5 shows the precision, recall, F1 score, and F1-weighted score of the proposed model and the baseline models on the test set for the second output, which is the urgent or not urgent classification. As shown in Table 5, the CBiGRU model also achieved the highest values compared to baseline models while classifying the urgency of the questions. CBiGRU model outperformed the second-best model (Bi-LSTM) by **0.6%** in the F1 score for the urgent class and **1.1%** in the F1-weighted score for the whole classification task. This demonstrated the effectiveness of combining CNN and Bi-GRU layers to capture both local and global features of the input data.

Table 4: Experimental results on question classification task

Model	Non-Question			Question			Weight F1 (%)
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
CNN	90.6	94.9	92.7	89.5	81.6	85.4	90.1
LSTM	95.3	95.6	95.4	92	91.1	91.5	94
Bi-LSTM	95.1	95.7	95.3	92.1	91.3	91.6	94.1
GRU	95	95.5	95.2	92.1	90	91	93.7
Bi-GRU	95.4	95.6	95.4	92.5	**91.6**	92	94.3
CNN + Bi-LSTM	94.9	**96.6**	95.7	93.3	90.3	91.7	94.3
CNN + Bi-GRU (CBiGRU)	**95.8**	96.4	**96**	**93.5**	91.3	**92.3**	**94.8**

Table 5: Experimental results on urgency classification task

Model	Non-Urgent			Urgent			Weight F1 (%)
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
CNN	87.3	86.1	86.6	84.2	85.3	84.7	85.7
LSTM	**88.9**	86.1	87.4	84.2	87.3	85.7	86.6
Bi-LSTM	86.8	89.4	88	85	87.8	86.3	87.4
GRU	86.6	86.8	86.7	84.6	84.4	84.5	85.7
Bi-GRU	87.6	85.9	86.7	83.4	86.3	84.8	85.8
CNN + Bi-LSTM	86.7	90.4	88.5	**85.2**	87.1	86.1	87.2
CNN + Bi-GRU (CBiGRU)	88.4	**91.7**	**90**	84.2	**89.8**	**86.9**	**88.5**

Precision-recall curves show the trade-off between precision and recall for different threshold values. They can help us to compare the performance of different models across the range of possible classifications. The area under the curve (AUC) is a measure of how well the model can distinguish between the classes. Figures 15 to 17 show the precision-recall curves for baseline models and the proposed CBiGRU model. As we can see from the precision-recall curves, the CBiGRU model achieved a wider precision-recall coverage area compared to the baseline models for both outputs, which indicates that it can achieve high precision and recall for any threshold value. CBiGRU model also has the highest AUC values of **93.6** and **89.7** for the first and second outputs, respectively, while the second-best

model (CNN + BiLSTM) has AUC values of **93.3** and **88.6** for the first and second outputs, respectively. This shows that the proposed model has a stronger ability to separate the classes than the baseline models for both outputs.

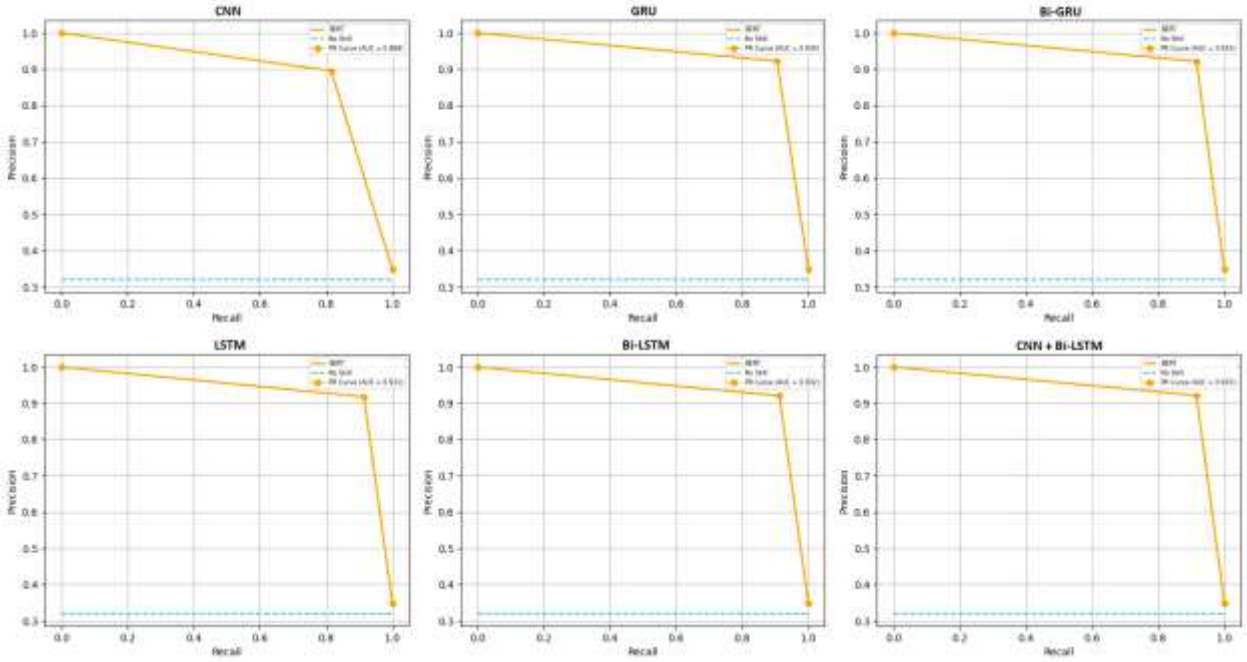

Figure 15: PR curves for CNN, LSTM, GRU, Bi-LSTM, Bi-GRU and CNN + Bi-LSTM models for question classification task

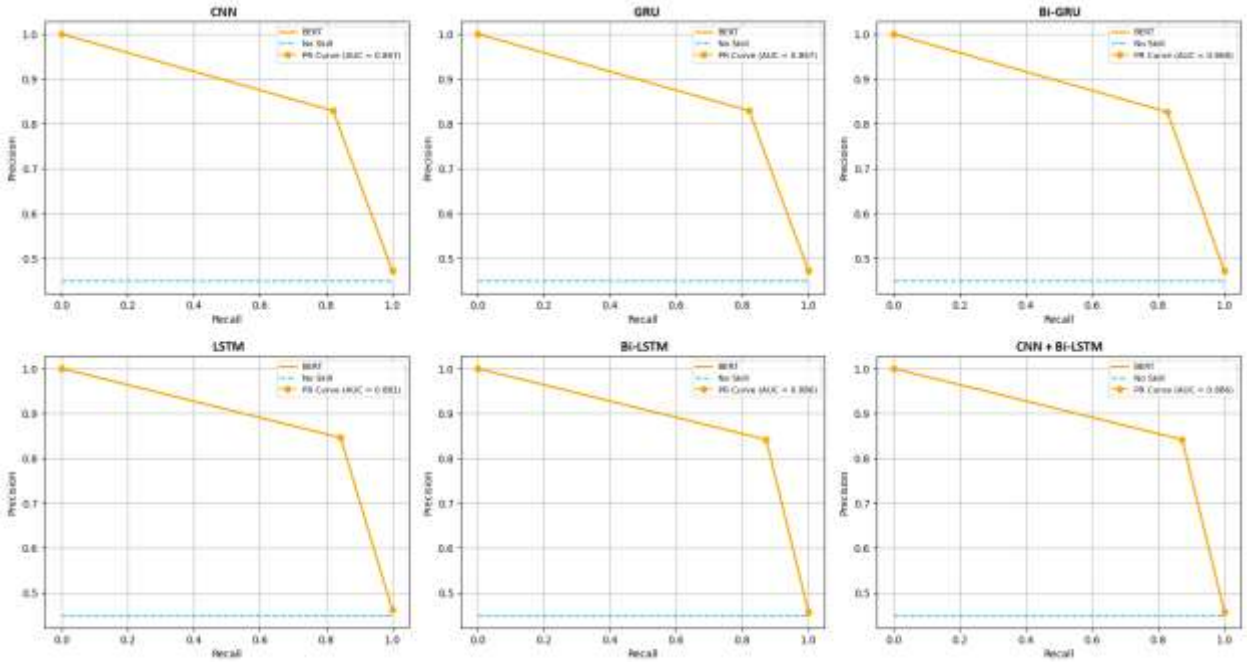

Figure 16: PR curves for CNN, LSTM, GRU, Bi-LSTM, Bi-GRU and CNN + Bi-LSTM models for question classification task

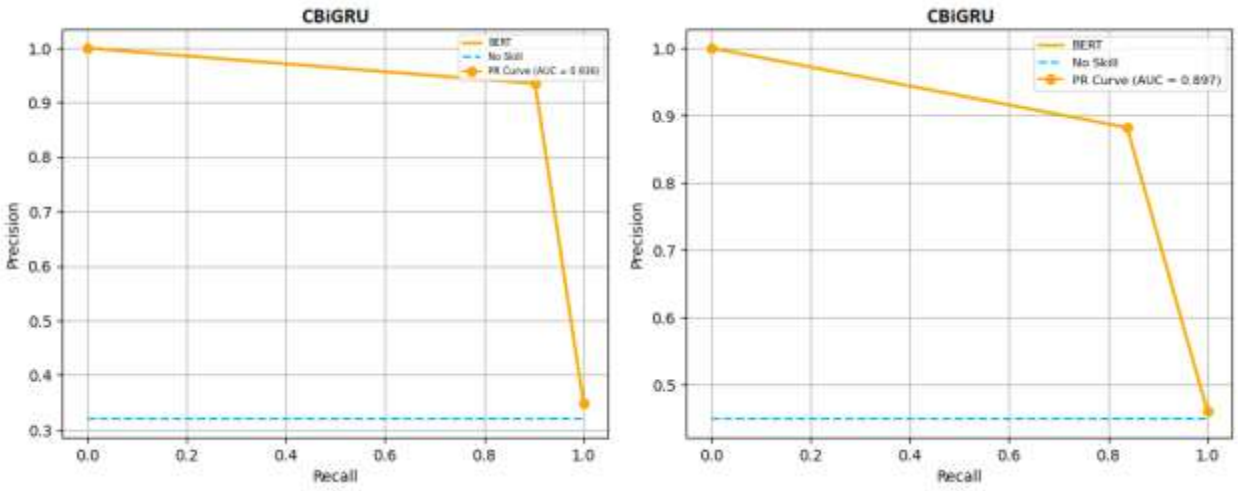

Figure 17: PR curves for proposed CNN + Bi-GRU model named CBiGRU for both classification tasks

To explore the impact of language formality on model performance, we evaluated four pre-trained BERT models: BERT-base-cased, BERT-base-uncased, BERT-large-cased, and BERT-large-uncased. As presented in Table 6, BERT-base-uncased achieved better performance compared to the other models. This finding suggests that MOOC discussion forums, often characterized by informal language and spelling errors, might benefit from case-insensitive approaches. Thus, the simplicity of BERT-base-uncased, which ignores letter case, proved to be more effective in this context.

To assess the influence of dataset balance, we evaluated the model's performance on both imbalanced and balanced datasets. As shown in Table 7, employing BERT augmentation to balance the data led to a statistically significant improvement in model performance. This finding aligns with the established principle that balanced datasets can enhance the effectiveness of deep learning models for binary classification tasks.

We further conducted a series of experiments varying key parameters within convolutional and max-pooling layers to understand the relationship between CNN architecture and model performance. This exploration aimed to identify the configuration that optimizes performance for our specific task. The influence of these alterations on the model's performance is outlined in Table 8.

Table 7: Comparative analysis of pretrained BERT model.

Model	Question Classification							Urgency Classification						
	Non-Question			Question			Weight F1 (%)							Weight F1 (%)
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
BERT-base-cased	95.2	93.1	94.1	88.8	91.6	90.2	92.3	88.1	87.9	87.9	80.5	88.2	84.1	86.2
BERT-base-uncased	95.8	96.4	96	93.5	91.3	92.3	**94.8**	88.4	91.7	90	84.2	89.8	86.9	**88.5**
BERT-large-cased	96.3	91.1	93.6	86.3	89.5	87.8	90.9	88.3	86.1	87.1	82.7	85.2	84	85.6
BERT-large-uncased	95.3	92.8	94	90.1	90.8	90.4	92.4	86.5	87.2	86.8	83.1	87	86.9	86.8

Table 6: Comparative analysis of model performance on balanced and imbalanced datasets

Model	Question Classification							Urgency Classification						
	Non-Question			Question			Weight F1 (%)							Weight F1 (%)
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
Imbalanced	92.9	95.2	94	**94.4**	89.1	91.6	92.9	87.1	89.3	88.1	82.7	88.6	85.5	86.9
Balanced	**95.8**	**96.4**	**96**	93.5	**91.3**	**92.3**	**94.8**	**88.4**	**91.7**	**90**	**84.2**	**89.8**	**86.9**	**88.5**

Table 8: Correlation Between CNN Parameters and Model Performance

CNN Kernel	Number of CNN filters	Pool Size (Maxpooling)	Question Classification							Urgency Classification						
			Non-Question			Question			Weight F1 (%)							Weight F1 (%)
			P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	
CNN – 4	64	8	92.1	92.8	92.4	89.3	88.6	88.9	90.8	84.8	85.6	85.1	82.5	84.9	83.6	84.4
CNN – 5	64	8	92.7	93	92.8	91	90.9	90.9	91.9	85.4	86.1	85.7	83.2	85.1	84.1	85
CNN – 6	64	8	92.6	93.1	92.8	90.3	90.7	90.4	91.7	86.3	87.1	86.6	82	86.8	84.3	85.6
CNN – 7	64	8	91.7	93.3	92.4	91.7	88.4	90	91.3	85	87.8	86.3	81.6	86.4	83.9	85.2
CNN – 4	128	8	93.9	94.4	94.1	91.5	90.2	90.8	92.6	84.7	89.5	87	82.4	88.1	85.1	86.1
CNN – 5	128	8	93.1	95.3	94.2	92.7	92.4	92.5	93.4	87.1	88.9	88	83.2	87.6	85.3	86.7
CNN – 6	128	8	92.5	93.9	93.1	91.8	91.1	91.4	92.3	87.4	89.3	88.3	81.8	88.2	84.8	86.7
CNN – 7	128	8	92.1	93.6	92.8	92.3	88.8	90.5	91.7	85.8	87.9	86.8	82.7	87.5	85	86
CNN – 4	256	8	94.3	94.8	94.5	91.2	90	90.6	92.7	86.1	88.9	87.4	83.1	86.2	84.6	86.1
CNN – 5	256	8	92.8	94.6	93.6	92.6	89.2	90.8	92.3	87.3	89.6	88.4	84.5	86.7	85.5	87.1
CNN – 6	256	8	93.6	94.3	93.4	91.7	90.8	91.2	92.4	83.5	91.3	87.2	83.1	88.1	85.5	86.4
CNN – 7	256	8	91.1	94.2	92.6	91.8	88.4	90	91.4	82.9	90.9	86.7	82.5	86.4	84.4	85.6
CNN – 3 – 4	64 – 64	8 – 4	93.8	94.7	94.2	92.1	90.8	91.4	92.9	86.7	89.4	88	82.7	88.2	85.3	86.7
CNN – 3 – 5	64 – 64	8 – 4	94.1	94.9	94.4	92.7	91	91.8	93.2	85.2	91.1	88.1	83.8	86.3	85	86.6
CNN – 3 – 6	64 – 64	8 – 4	95.3	96	95.6	93.1	90.3	91.6	93.8	87.9	90.3	89.1	84.1	88.1	86	87.6
CNN – 3 – 7	64 – 64	8 – 4	94.9	94.6	94.7	93.6	90.2	91.8	93.4	87.7	91.2	89.4	83.5	88.3	85.8	87.7
CNN – 3 – 4	128 – 128	8 – 4	95.1	95.3	95.1	92.1	91.9	91.9	93.7	88.1	90.8	89.4	83.2	87.7	85.3	87.5
CNN – 3 – 5	128 – 128	8 – 4	94.8	95.1	94.9	93.7	90.4	92	93.5	88	90.6	89.2	84.8	88.3	86.5	88
CNN – 3 – 6	128 – 128	8 – 4	95.8	96.4	96	93.5	91.3	92.3	**94.8**	88.4	91.7	90	84.2	89.8	86.9	**88.5**
CNN – 3 – 7	128 – 128	8 – 4	95.2	95.7	95.4	92.8	91.1	91.3	93.8	89.2	90.1	89.6	83	89.2	85.9	87.9
CNN – 3 – 4	256 – 256	8 – 4	93.9	95.1	94.4	93.1	90	91.5	93.1	87.8	89.4	88.5	84.1	88.7	86.3	87.5
CNN – 5 – 5	256 – 256	8 – 4	94.5	95.6	95	92.8	90.4	91.5	93.4	88.3	88.6	88.4	83.3	86.5	84.8	86.7
CNN – 3 – 6	256 – 256	8 – 4	94.2	95.7	94.9	93	91.1	92	93.6	86.4	87.2	86.7	82.1	85.4	83.7	85.3
CNN – 3 – 7	256 – 256	8 – 4	93.1	94.3	93.6	92.7	89.4	91	92.4	83.9	88.3	86	83	86.1	84.5	85.3
CNN – 3 – 4 – 5	64 – 64 – 64	8 – 4 – 2	92.7	94.1	93.3	91.8	90.2	90.9	92.3	85.4	86.6	85.9	81.5	83.7	82.5	84.4
CNN – 4 – 5 – 6	64 – 64 – 64	8 – 4 – 2	92.1	93.7	92.8	90.3	89.7	90	91.5	85	87.1	86	80.6	82.3	81.4	83.9
CNN – 5 – 6 – 7	64 – 64 – 64	8 – 4 – 2	91.7	93.8	92.7	91.7	89.3	90.4	91.6	83.8	86.3	85	81.2	82.8	81.9	83.6
CNN – 3 – 4 – 5	128 – 128 – 128	8 – 4 – 2	93.3	95.2	94.2	92.4	90.7	91.5	92.9	86.1	87.4	86.7	83.1	81.2	82.1	84.6
CNN – 4 – 5 – 6	128 – 128 – 128	8 – 4 – 2	93.7	94.7	94.1	92.9	91.3	92.1	93.2	87	86.8	86.8	82.5	84.6	83.5	85.3
CNN – 5 – 6 – 7	128 – 128 – 128	8 – 4 – 2	93.1	94.2	93.6	91.8	91	91	92.6	86.7	86.2	86.4	83.1	85.8	84.4	85.5
CNN – 3 – 4 – 5	256 – 256 – 256	8 – 4 – 2	92.8	94.7	93.7	91	90.7	90.8	92.4	86.1	86.9	86.4	82.8	84.6	83.6	85.2
CNN – 4 – 5 – 6	256 – 256 – 256	8 – 4 – 2	93.3	93.8	93.5	92.3	91.8	91	92.8	87.3	87.9	87.5	82.3	83.1	82.6	85.3
CNN – 5 – 6 – 7	256 – 256 – 256	8 – 4 – 2	92.7	93	92.8	91.4	90.3	90.8	91.9	85.4	87.1	86.2	81.9	84.5	83.1	84.8

6. CONCLUSION

In the rapidly evolving landscape of Massive Open Online Courses (MOOCs), extracting and classifying urgent student questions emerge as a critical yet understudied facet. In this paper, we proposed a novel hybrid CNN + BiGRU multi-output classification model, which can automatically identify and prioritize student questions from online discussion forums. We addressed the issue of extracting and classifying urgent student questions, which has not been explored in the previous literature. Our model leverages the advantages of both CNN and BiGRU layers to capture both local and global features of the input data. We employed the BERT pre-trained model for word embedding and balancing the dataset, which are two key factors that improved the performance of our model. We showed that the BERT model can provide rich and contextualized representations of the input data, and balancing the dataset can reduce the bias and the noise of the data.

We evaluated the proposed model on a manually labeled dataset collected from 11 public online courses at Stanford University and compared it with six baseline models. The results showed that our model outperformed the baseline models in most cases for both outputs. CBiGRU model achieved high accuracy and coverage in identifying and prioritizing student questions, which can help instructors and tutors provide timely and appropriate feedback to the students.

The model relies on the textual features of the input data and does not consider other features, such as the metadata of the posts, the user profiles, the ratings, the comments, or the temporal information. Future work may consider these features which might provide useful information for the classification task, and improve the model performance. As well as, the direction for future work is to focus on question visualization and aspect-based question extraction, which can provide more insights and details about the student question data, and help instructors and tutors to better understand and respond to the student's needs and interests.

Acknowledgement

This work was supported by Shahid Rajaee Teacher Training University.

Funding

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

Availability of data and material

Data are available on justified request to the corresponding author.

Credit author statement

Mujtaba Sultani: Conceptualization, Data curation, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft.

Negin Daneshpour: Conceptualization, Methodology, Writing – review & editing.

REFERENCES

[1] Paufler, N. A., Sloat, E. F.: Using standards to evaluate accountability policy in context: School administrator and teacher perceptions of a teacher evaluation system. Studies in Educational Evaluation, 64, 2020, 100806. https://doi.org/10.1016/j.stueduc.2019.07.007

[2] Chaturvedi, S., Goldwasser, D., Daumé III, H.: Predicting instructor's intervention in mooc forums. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 1501-1511.

[3] Stephens-Martinez, K., Hearst, M. A., Fox, A.: Monitoring moocs: which information sources do instructors value?. In Proceedings of the first ACM conference on Learning@ scale conference, 2014, pp. 79-88. https://doi.org/10.1145/2556325.2566246

[4] Zhang, C., Chen, H., Phang, C. W.: Role of instructors' forum interactions with students in promoting MOOC continuance. Journal of Global Information Management (JGIM), 26(3), 2018, 105-120. https://dOI: 10.4018/JGIM.2018070108

[5] Pena, W., Melgar, A.: Ontology-based information extraction from Spanish Forum. In Computational Collective Intelligence: 7th International Conference, ICCCI 2015, Madrid, Spain, September 21-23, 2015, Proceedings, Part I (pp. 351-360). Springer International Publishing. https://doi.org/10.1007/978-3-319-24069-5_33

[6] Kizilcec, R. F., Piech, C., Schneider, E.: Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In Proceedings of the third international conference on learning analytics and knowledge, 2013, pp. 170-179. https://doi.org/10.1145/2460296.2460330

[7] McDonald, J., Moskal, A. C. M., Goodchild, A., Stein, S., Terry, S.: Advancing text-analysis to tap into the student voice: A proof-of-concept study. Assessment & Evaluation in Higher Education, 45, 2020, 154–164. https://doi.org/10.1080/02602938.2019.1614524

[8] Bakharia, A.: Towards cross-domain MOOC forum post classification. In *Proceedings of the third (2016) ACM conference on learning@ scale* (pp. 253-256). https://doi.org/10.1145/2876034.2893427

[9] Wei, X., Lin, H., Yang, L., Yu, Y.: A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information*, 8(3), 2017, 92. https://doi.org/10.3390/info8030092

[10] Almatrafi, O., Johri, A., Rangwala, H.: Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. *Computers & Education, 118, 2018*, 1-9. https://doi.org/10.1016/j.compedu.2017.11.002

[11] Guo, Z. X., Sun, X., Wang, S. X., Gao, Y., Feng, J.: Attention-based character-word hybrid neural networks with semantic and structural information for identifying of urgent posts in MOOC

discussion forums. *IEEE access*, 7, 2019, 120522-120532. https://doi.org/10.1109/ACCESS.2019.2929211

[12] Khodeir, N. A.: Bi-GRU urgent classification for MOOC discussion forums based on BERT. *IEEE Access*, 9, 2021, 58243-58255. https://doi.or/ 10.1109/ACCESS.2021.3072734

[13] El-Rashidy, M. A., Farouk, A., El-Fishawy, N. A., Aslan, H. K., Khodeir, N. A.: New weighted BERT features and multi-CNN models to enhance the performance of MOOC posts classification. *Neural Computing and Applications*, 2023, 1-15. https://doi.org/10.1007/s00521-023-08673-z

[14] Almatrafi, O., Johri, A.: Improving MOOCs Using Information From Discussion Forums: An Opinion Summarization and Suggestion Mining Approach. IEEE Access, 10, 2022, 15565-15573. https://doi.org/ 10.1109/ACCESS.2022.3149271

[15] Talebi, K., Torabi, Z., Daneshpour, N.: Ensemble models based on CNN and LSTM for dropout prediction in MOOC. *Expert Systems with Applications*, *235*, 2024, 121187. https://doi.org/10.1016/j.eswa.2023.121187

[16] Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado-Kloos, C., Munoz-Organero, M.: Delving into participants' profiles and use of social tools in MOOCs. *IEEE Transactions on Learning Technologies*, *7*(3), 2014, 260-266. https://doi.org/ 10.1109/TLT.2014.2311807

[17] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., Seaton, D. T.: Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment*, *8*, 2013, 13-25.

[18] Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., Sherer, J.: Social factors that contribute to attrition in MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference*, *2014*, pp. 197-198. https://doi.org/10.1145/2556325.2567879

[19] Feng, L., Liu, G., Luo, S., Liu, S.: 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. https://doi.org/10.1007/978-3-319-70093-9_39

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

[21] Macina, J., Srba, I., Williams, J. J., Bielikova, M.: Educational question routing in online student communities. In *Proceedings of the eleventh ACM conference on recommender systems*, *2017*, pp. 47-55. https://doi.org/10.1145/3109859.3109886

[22] Cui, Y., Wise, A. F.: Identifying content-related threads in MOOC discussion forums. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 299-303). https://doi.org/10.1145/2724660.2728679

[23] Agrawal, A., Paepcke, a.: The stanford MOOC Posts dataset, 2014. Accessed: Dec, 15, 2020.

[24] Shaikh, S., Daudpota, S. M., Imran, A. S., Kastrati, Z.: Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. Applied Sciences, 11(2), 2021, 869. https://doi.org/10.3390/app11020869

[25] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 2002 321-357. https://doi.org/10.1613/jair.953

[26] He, H., Bai, Y., Garcia, E. A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), 2008, pp. 1322-1328. Ieee. https://doi.org/10.1109/IJCNN.2008.4633969

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

[28] Rao, D., McMahan, B.: *Natural language processing with PyTorch: build intelligent language applications using deep learning*. " O'Reilly Media, Inc.", 2019.

[29] Harris, Z. S.: Distributional Structure, WORD, vol. 10, 1954 no. 2–3.

[30] Wang, S., Zhou, W., Jiang, C.: A survey of word embeddings based on deep learning. *Computing*, *102*, 2020, 717-740. https://doi.org/10.1007/s00607-019-00768-7

[31] Mikolov, T., Yih, W. T., Zweig, G.: Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746-751).

[32] Yoav, H., Goldberg, Graeme: *Neural Network Methods in Natural Language Processing*. San Rafael, CA, USA: Morgan & Claypool, 2017, 104–113.

[33] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 2017, 135-146. https://doi.org/10.1162/tacl_a_00051

[34] Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

[35] Luan, Y., Lin, S.: Research on text classification based on CNN and LSTM. In 2019 IEEE international conference on artificial intelligence and computer applications (ICAICA), 2019, pp. 352-355. IEEE. https://doi.org/ 10.1109/ICAICA.2019.8873454

[36] Du, J., Vong, C. M., Chen, C. P.: 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), 2020, 1586-1597. https://doi.org/10.1109/TCYB.2020.2969705

[37] Schuster, M., Paliwal, K. K.: Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 1997, 2673-2681. https://doi.org/ 10.1109/78.650093

[38] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014. *arXiv preprint arXiv:1406.1078*. https://doi.org/10.48550/arXiv.1406.1078

[39] Lu, Q., Zhu, Z., Xu, F., Zhang, D., Wu, W., Guo, Q.: Bi-gru sentiment classification for chinese based on grammar rules and bert. *International Journal of Computational Intelligence Systems*, *13*(1), 2020, 538-548. https://doi.org/10.2991/ijcis.d.200423.001

[40] Anzanello, M. J., Fogliatto, F. S.: Learning curve models and applications: Literature review and research directions. International Journal of Industrial Ergonomics, 41(5), 2011, 573-583. https://doi.org/10.1016/j.ergon.2011.05.001