I. 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 the 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 result 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 researches.

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, none have focused specific 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. Recognizing this, the efficient extraction of urgent questions becomes even more critical, ensuring prompt responses to address learners' immediate concerns [8]. The need of 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 [9]. Previous studies highlighted the correlation between active forum participation, successful question answering, and reduced dropout rates [10 - 12]. This underscores the importance of extracting urgent questions and providing timely answers to student queries which can boost retention, engagement, and, importantly, help decrease the dropout rates. We propose 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 the combination of Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU) layers. In 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. We propose CBiGRU, a BERT-based multi-output hybrid deep learning model to extract urgent student questions from MOOC discussion forums.
2. The first research paper in the education domain, particularly within the context of MOOCs, to extract and classify urgent student questions.
3. Implementation of BERT-based data augmentation technique to address dataset imbalance and enhance the model's ability to generalize across diverse educational contexts.
4. Employing BERT for contextualized embedding which capture the contextual meaning of words within the MOOC forum discussions.
5. Helping instructors, course designers, and policymakers to enhance decisions on course participation, and improvement, with a focus on urgent student concerns.

II. 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 into urgent post classification within MOOCs, each contributing unique insights. Almatrafi and his co-authors [13] addressed this challenge by utilizing metadata, linguistic features, and traditional machine learning algorithms, with AdaBoost produced a better result. Feng et al. [14] 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 impressive overall 85% accuracy. Wei and his colleagues [15] employed a convolutional LSTM-based deep neural network to classify confusion, urgency, and sentiment in MOOC discussion forums. They learned word-level features with convolutions 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. [16] 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 [17] proposed a classification model to identify confusion and recommend optimal start times for video clips. The paper leveraged features like bag-of-words, post metadata, and made predictions across various labels. Furthermore, Khodeir [18] utilized BERT for embedding and Bi-GRU as classification algorithm and obtained 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. In a recent study, El-Rashidy and his co-authors [19] presented a four-stage model which 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 [20] 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.

III. 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 which apply two-step classification process. Initially, the model identifies whether a student 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 proposed multi-output classification approach. Figure 1 represents the sequential flow of these operations, providing a concise overview of the proposed approach.

a. Preprocessing

In the realm of applying Deep Learning to text, several preprocessing steps are necessary to ensure meaningful analysis. 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 which maintains semantic integrity.
2. **URLs Removal:** URLs are removed to eliminates 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 [16], to significantly improve overall results.

b. 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] [25] [26] [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 on 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].

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 2 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.

c. 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.

i. 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 [18]. 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 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 Neural probabilistic model, Word2Vec and FastText. Neural probabilistic model learns embeddings by predicting a word's context through one-hot encoded vectors and back-propagation. Word2Vec 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, 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 [18].

In addition to prediction-based models, the BERT introduces context-aware word representations and allow 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. As shown in Figure 3, 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 [18]. 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.

ii. C**BiGRU 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 gradient 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 bidirectionally. One layer processes the input sequence from the first-time step to the last, while the other processes it in reverse, from the last-time 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 comparing 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.

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].

As shown in Figure 5, 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. These embedding then pass through one-dimensional convolutional layer with 128 filters and a kernel size of 3, followed by a max-pooling layer to capture local features. To prevent overfitting, a dropout layer with a rate of 0.2 is added. 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 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.

IV. EXPERIMENT

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

a. Dataset

In this paper, the experiments are conducted using the Stanford MOOC Posts dataset, a standardized benchmark corpus introduced by Agrawal et al. (2015). The dataset contains 29,604 anonymized forum posts from 11 public online courses from Stanford University, stratified across three domains named Humanities/Sciences, Medicine, and Education. As shown in Figure 5, 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.

b. 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 pretrained ‘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.

c. 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 [18]. 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.

$$PR = TP / (TP + FP) \tag{6}$$

$$RC = TP / (TP + TN) \tag{7}$$

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

TP are correct positive predictions, TN (True Negatives) are correct negative predictions, and FP (False Positives) are incorrect positive predictions. 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 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 nuance insights into the model's performance, especially in imbalanced scenarios [18]. 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.

V. RESULT AND DISCUSSION

The proposed multi-output classification approach presented in this paper, designed to extracts and classifies urgent student questions, marks a novel contribution to the field. Therefore, directly comparison of our proposed model with previous studies may not be conceptually sound due to the different goals of this paper. However, concerning binary classification of student comments in MOOC discussion forums, our proposed models achieved better results compared to previous studies. As seen in Table 3, recent studies achieved F1-weighted score of 92.7% for the binary urgency classification task. 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 classifying the urgency level of student questions, our model achieved F1-weighted score of 88%, as shown in Table 5.

We compared our hybrid CBiGRU model with six baseline models: CNN, LSTM, BiLSTM, CNN+BiLSTM, GRU and BiGRU. As mentioned, learning curve can help us understand how well the model is learning from the data and whether it is overfitting or underfitting. Figure 10 to 16 show the LC 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.

Table 4 shows the precision, recall, F1-score and F1-weighted score of 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, CBiGRU model achieved better performance among all baseline models which indicate that it can correctly identify the posts that are questions with high confidence and coverage. The proposed model outperformed the second-best model (CNN + BiLSTM) by **0.04** in F1-score for the question class and **0.05** in 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, CBiGRU model also achieved the highest values comparing to baseline models while classifying urgency of the questions. CBiGRU model outperformed the second-best model (CNN + BiLSTM) by **0.06** in F1-score for the urgent class and **0.07** in F1-weighted score for the whole classification task. This demonstrated the effectiveness of combining CNN and BiGRU layers to capture both local and global features of the input data.