IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 29, No. 5, October, 2024, pp. 1891-1901
Journal homepage: http://iieta.org/journals/isi

Multi-Task Learning with BERT, RoBERTa, GPT-3.5, ELECTRA, and XLNet for Urgency Classification, Topic Similarity, and Sentiment Analysis in MOOCs

Aicha Marrhich[1*], Ichrak Lafram[2], Naoual Berbiche[1]

[1] Laboratory of Systems Analysis and Image Processing and Integrated Management, Higher Technology of Salé, Salé 11000, Morocco
[2] High Commisionner for planning, National Institue of Statistics and Applied Economics, Mohammed V University Agdal, Rabat 10080, Morocco

Corresponding Author Email: aichamarrhich@researchemi.ac.ma

https://doi.org/10.18280/isi.290521

Received: 23 March 2024
Revised: 28 June 2024
Accepted: 1 August 2024
Available online: 24 October 2024

Keywords:
MOOCs, deep learning, sentiment analysis, Large Language Models (LLMs), forums posts analysis, topic similarity, multi task models

ABSTRACT

In online education settings, effective student-teacher interaction can be challenging, often leaving student feedback on forums overlooked. This paper highlights the significance of student feedback and investigates the experiences shared on Massive Open Online Courses (MOOCs) platforms through forum interactions. We combined three distinct datasets from MOOCs forums into a unified corpus suitable for training Large Language Models (LLMs) for diverse classification tasks. Our study uses Bidirectional Encoder Representations from Transformers (BERT), a Robustly Optimized BERT Pretraining Approach (RoBERTa), Generative Pre-trained Transformer 3.5 (GPT -3.5), Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), and eXtreme Multi-head Attention Network (XLNet) to classify feedback based on urgency levels, assess topic similarity, and analyze sentiment to gauge overall classroom sentiment. The resulting multi-task learning framework addresses the classification of questions, urgency levels, and sentiment analysis concurrently, enhancing the management of student inquiries and satisfaction in MOOC environments. This research contributes methodologically by demonstrating the efficacy of LLMs in handling multifaceted feedback analysis tasks, thereby enriching the understanding of student engagement and satisfaction in online courses.

1. INTRODUCTION

Our previous work aimed at investigating how machine learning can enhance the learning experience in virtual classrooms [1] from the teacher's point of view. We listed some of the challenges that teachers face in online settings as well as the different roles they play that range from discussion and technology facilitator, social supporter, to course designer and differentiator. We emphasized the importance of real time feedback in forum discussions. Several tracks have emerged and converged into the necessity of providing a learning setting that is similar to a face to face one in order to guarantee that the components of success are present, which are: student instructor interaction, instant feedback, course design and assessment.

1.1 Context of the study

1.1.1 Learners' feedback serves as a key component in fuelling a successful learning experience

Educators in virtual settings are faced with so many challenges [2]. In order to assist students in virtual learning environments, teachers should monitor forum discussions, which can be a real tank for problems students usually face, since forums represent a way of expressing all sorts of issues that the learners might encounter throughout the course. However, it remains practically impossible for a teacher to tackle every student's comment and question because of the immense volume of questions. Therefore, assisting teachers in virtual settings can help tremendously to create a learning environment similar to a face-to-face setting in a way that a teacher is always there to provide relevant feedback to his students when needed.

In fact, feedback from forum discussions plays a crucial role in the learning journey of student (Figure 1). It can also be very useful in enhancing the learning experience from the teacher's point of view in 2 major ways which are:

-Course design and delivery
-Course assessment

Course design and delivery

The design and delivery of an online class can be challenging as it involves a great deal of issues such as which content should be prioritized, which notions would be easily retained and understood. Instructors should organize the course material into chunks as well as provide good theoretical explanations combined with opportunities for practice. The level of difficulty should be ascendent so as to allow students to build knowledge in a smooth way. With hindsight, teachers

use students 'interaction with the course content to adapt the material used to the students' needs and so decide about the relevance of the content regarding a given audience.

Course assessment

Assessment of knowledge is an integral step of teaching because it allows teachers to test if the education objectives have been met or not. Assessment can play a determinant role in adjusting and tailoring the course in a more efficient way, it can give great insights on the course methodology because it mirrors the areas student still need to make progress on. Also, it gives valuable information to instructors to reflect on how they might improve the courses in terms of materials, curriculum, methodology. Gibbs [3] states that one of the most important roles of assessment is to enable teachers to decide of the appropriateness of standards on the course, Gibbs also evaluates the impact of assessment on student learning and concludes its effectiveness in supporting learning.

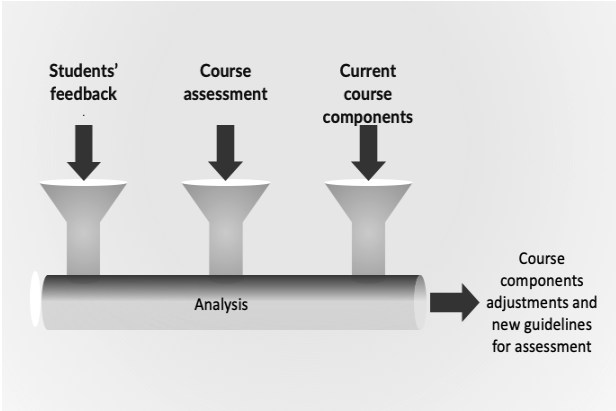

Figure 1. Students' feedback feeding into course design and course assessment

Students' feedback

Hattie and Timperley [4] agreed that what makes the biggest difference in student performance is feedback. Still, another challenge that arises with virtual settings is the lack of instructor-teacher feedback that serves both the instructor and the student to make the learning experience more efficient. From one hand, teachers are unable to track students' understanding of the components of the lesson. On the other hand, students don't get their questions answered properly and may feel confused and demotivated if they are struggling with some key concepts to push their learning journey a step forward.

Overall, Students' feedback stands as the key component that is injected into course design and course assessment in order to enhance the teaching learning experience. With students' feedback, teachers can design better courses and assess students accordingly.

1.1.2 The wheel of continuous teaching improvement that feeds on students' feedback analysis

The wheel of continuous teaching improvement, when implemented properly (Figure 2), can prove to enhance the learning experience through the following steps:

-Identifying educational goals

-Identifying specific knowledge, skills to meet each goal

-Organizing knowledge from the most basic to the most advanced

-Designing learning activities and assessment

-Evaluating the learning experience

In light of the above, we can see that teachers can really design effective and engaging learning experiences that calls on knowledge and skills needed to succeed.

In order to incorporate this wheel of continuous development in online settings, teachers need to rely on students' feedback in the form of questions, opinions and inquiries to feed their wheel with inputs about the overall grasp of the course content. Students' feedback allows the teachers to figure out technical difficulties and lack of structure in their courses, to consider the variety of students learning styles and to diversify course content.

Also, feedback analysis can prove to be helpful in creating tests that progress in difficulty and complexity, tests become challenging and can measure learners' knowledge acquisition more accurately (Figure 3). The tests outcomes can help improve future assessment tools in a way that provides more balanced and comprehensive tests that gradually broaden the learning spectrum and cover skills ranging from approachable to more complex.

In light of these considerations, our study focuses on utilizing advanced language models such as Bert, RoBERTa, GPT-3.5, ELECTRA, and XLNet to analyze student feedback in MOOCs forums. Specifically, we aim to classify feedback based on urgency levels, assess topic similarity, and analyze sentiment to gauge classroom sentiment effectively. By implementing a multi-task learning framework, we seek to enhance educators' ability to manage student inquiries and satisfaction in online courses more efficiently.

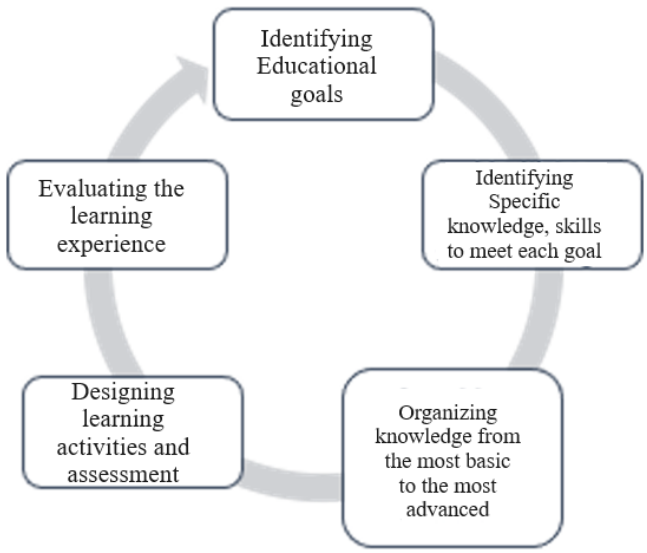

Figure 2. The wheel of continuous course improvement

Figure 3. The importance of students' feedback in course design and assessment

2. LITERATURE REVIEW

2.1 Existing work on classifying feedback in MOOCs

In our research, we examined previous work on classifying

feedback, we grouped the findings into three categories: content analysis, sentiment analysis and natural language processing.

Content analysis

A leading study recommends utilizing BERT, a contextual word representation model, for classifying posts and identifying urgent ones within the Stanford MOOC posts dataset. The model's output is then fed into a multi-layer bi-directional Gated Recurrent Unit (GRU) for further processing [5].

Recent research advocates the use of deep learning techniques for classifying student feedback. It examines various methods, such as a convolutional Long Short-Term Memory (LSTM) network for classifying MOOC forum posts across different domains. This approach aims to categorize feedback, aiding educators in addressing key student concerns. The study suggests that this model could enhance real-time monitoring of MOOC forums [6].

Another research proposes a technique for topic modeling to categorize discussion threads and then conducts sentiment analysis to assess the emotional tone of posts. This method is intended to assist instructors in managing MOOCs more effectively [7].

Similarly, a significant study analyzed extensive open-ended feedback from MOOC learners using LDA Topic Modeling and qualitative analysis, demonstrating that feedback can notably influence the learning experience [8].

Another analysis focused on discussions from forums across 60 MOOCs offered by Coursera, categorizing them based on interaction types [9].

In a related study, Atapattu et al. [10] classified emerging discussion topics to help identify influential clusters and highlight discussions requiring instructor intervention.

Sentiment analysis

A recent investigation explored student sentiment and opinions through sentiment analysis of comments posted in the forums of 60 MOOCs, studying the relationship between student sentiment and course ratings [11].

Another study analyzed collective sentiment from MOOC forum posts to track changing student opinions about the course, highlighting the link between sentiment ratios and daily dropout rates [12].

A further study employed sentiment analysis using Python and NVivo tools on feedback from three MOOCs—Introduction to Cybersecurity, Digital Forensics, and Development of Online Courses for SWAYAM. The findings revealed key factors affecting learner retention [13].

Natural language processing

Shaik et al. [14] discussed trends and challenges in applying NLP methods for analyzing educational feedback, focusing on techniques such as sentiment annotation, entity recognition, text summarization, and topic modeling for educational purposes.

Sun et al. [15] introduced a model designed to identify "urgent" posts needing immediate instructor attention using deep learning techniques, including enhanced recurrent convolutional neural networks. This model aims to help educators navigate discussion forums efficiently, supporting timely interventions and potentially reducing dropout rates.

Similarly, Kumar and Troussas [16] proposed a method using various dimensions of learner posts to determine the necessity for urgent intervention.

In light of the above, it can be inferred that feedback analysis plays a significant role and can potentially contribute to the enhancement of the learning experience by facilitating the teacher's role and providing a face to face like learning experience where instructors and learners interact on various topics. The next chapters will suggest a feedback analysis framework that aims to classify students' discussions in a dataset that resulted from merging three datasets into one, the Stanford dataset, online students Forum, and NUS MOOC dataset.

3. DATA CONSTRUCTION

Datasets to data science are indispensable because they feed in data to train and assess machine learning algorithms. Many data science techniques use algorithms and their performance depends on the quality and quantity of the data they are trained on. In fact, in order for an algorithm to make predictions and analyze data accurately, it needs to feed on good quality and large quantity of data.

In our case, these algorithms use large amounts of data to learn how to understand and process natural language. Data is crucial to train algorithms that enable machines to understand and process natural language. Though, it can be really difficult to find datasets for specific NLP tasks as these might vary widely. Also, their availability may change over time as new datasets are created and old ones removed or become outdated.

The construction of a contextual dataset to train machine learning models is challenging for many reasons, first, it can be quite hard to determine relevant context and select data pertaining to that context, second, the collection of data in a way that is usable for training machine learning models can be time-consuming and may require specialized skills, third, another challenge arise to ensure that the date is accurate and free of errors because this impacts directly the performance of the models that are trained on these datasets.

Chapter 1 For this study, we collected data from multiple MOOCS platforms 'forums to create a comprehensive dataset that reflects diverse student interactions. The forums were chosen based on their popularity and availability of student feedback across various courses and subjects. Each forum provided a rich source of textual data including student queries, comments, and discussions related to course content and learning experiences.

3.1 Existing datasets

To build our new dataset for the context of MOOCs, we started by exploiting existing datasets that are relevant to the topic. We carefully selected three datasets based on their quality, reliability and relevance to our research goals. Once we had identified the appropriate datasets, we began the process of cleaning and preprocessing the data in order to make it ready for analysis. This involved removing any irrelevant or redundant information, as well as ensuring that the data was consistent and well-structured. We also performed various other preprocessing steps, such as tokenization, stemming, and lemmatization, in order to prepare the data for further analysis.

The Stanford dataset (30000)

The Stanford MOOC Posts Dataset [17] encompasses 29,604 anonymized posts collected from learner forums across eleven publicly accessible online courses offered by Stanford University. This dataset is available for academic research upon request and was specifically created to support

computational studies of MOOC (Massive Open Online Course) discussions. The posts have been categorized into three distinct groups based on course subjects: Humanities/Sciences, Medicine, and Education, with each group containing 10,000, 10,002, and 10,000 posts respectively. The Humanities/Sciences group includes courses in economics, statistics, global health, and environmental physiology; the Medicine group features courses in medical statistics, science writing, and emergency medicine; and the Education group comprises a single course titled "How to Learn Math." Each post in the dataset is evaluated across six criteria:

Question: Identifies whether the post contains a question.

Opinion: Determines if the post presents an opinion or if it is purely factual.

Answer: Assesses if the post serves as a response to a query from another learner.

Sentiment: Evaluates the post's sentiment on a scale from 1 (highly negative) to 7 (highly positive), with 4 representing a neutral sentiment.

Urgency: Measures the urgency for an instructor's response on a scale from 1 (not urgent) to 7 (highly urgent), where 4 suggests that a response is needed only if the instructor has extra time.

Confusion: Rates the level of confusion expressed in the post from 1 (expert-level knowledge) to 7 (extreme confusion), with 4 indicating a state of neither confusion nor expertise.

Online student's forum (2000)

This dataset includes MOOCs discussion posts, it contains about 2057 threads and their corresponding tags which are as follows:

•O-discussion on subject theory, Q-Questions belonging to A category.

•B-Technical/software issue, BQ-Question belonging to B category.

•C-Logistics/deadline related discussion.

•S-Comments related to people socializing and introducing each other.

•P- showing politeness, e.g., saying thank you for appreciating something.

•T-Something that is completely off-topic and doesn't belong to any tag described earlier.

NUS MOOCs dataset (5000)

The NUS MOOCs dataset [18] is a comprehensive and multifaceted collection of discussion forum data from MOOCs, comprising a total of 33,665 threads across a diverse range of disciplines, including sciences, humanities, and engineering. This dataset was compiled from 61 completed courses on the Coursera platform, representing approximately 8% of Coursera's full course offerings. Coursera enables instructors to organize discussions by segmenting forums into various sub-forums, each dedicated to different aspects of the course or specific topics of focus. Generally, there is considerable variation in the number of instructor interventions across these sub-forums, with the most frequent interventions occurring in threads about exams and course logistics. Students are most active in sub-forums related to homework, quizzes/exams, and weekly lectures.

The dataset is composed of a multiclass, multi-level tagging system, which can be broken down into the following categories:

Request

•Ask for feedback

•Request justification

Elaborates

•Expansion

•Contrast

•Explanation

•Fine-tuning

•Evaluation of reasoning

Resolves

•Finalization

•Restatement

•Synthesis and summary

•Agreement

•Generic answer

•Disagreement

•Appreciation

•Other logistics

Social

•Social

•Other logistics

3.2 Data preprocessing

The collected textual data underwent rigorous preprocessing to ensure consistency and quality for subsequent analysis steps. The preprocessing pipeline included the following steps:

Text cleaning: We removed HTML tags, special characters, and non-textual elements that could distort analysis results.

Tokenization: Textual data was tokenized into individual words or sub words to facilitate further analysis by language models.

Normalization: Text normalization techniques such as lowercasing, stemming, and lemmatization were applied to standardize textual variations and reduce dimensionality.

Stop words removal: Common stop words that do not contribute significant meaning to the text were eliminated to focus on content-bearing words.

Handling missing data: Any incomplete or corrupted entries were identified and either corrected or removed to maintain data integrity.

3.3 The combined dataset

The process of combining three existing datasets involved implementing crucial cleaning and preprocessing steps. Each of the initial datasets possessed unique classes and information, which required us to devise an algorithm to harmoniously match and integrate these classes into a cohesive and homogeneous final dataset.

The initial phase involved meticulously cleaning and standardizing the individual datasets to ensure consistent formatting and quality. This entailed removing duplicate entries, handling missing values, and resolving any inconsistencies within the data. By carrying out these preprocessing steps, we focused on eliminating any potential biases or noise that could hinder the classification process.

Next, we worked on aligning the classes across the three datasets to establish a unified framework for classification (Figure 4). As the datasets originally contained distinct sets of classes, we developed an algorithm capable of mapping and matching similar classes to form a cohesive structure. This algorithm effectively identified analogous categories and merged them, creating a comprehensive and consolidated set

of classes for the final dataset.

The final dataset comprises three distinct classes: Question, Urgency, and Sentiment. These classes were carefully selected to capture essential aspects of language understanding and classification tasks.

The "Question" class encompasses instances where the text in the dataset represents queries or inquiries seeking information or clarification. This class is crucial for tasks involving question answering, information retrieval, and dialogue systems.

The "Urgency" class pertains to instances where the text expresses a sense of urgency or importance. This class enables the classification of texts based on their level of urgency, allowing for prioritization and timely response.

The "Sentiment" class captures the emotional tone or sentiment conveyed by the text. It allows for text classification into categories such as positive, negative, or neutral sentiments. This class is particularly valuable in sentiment analysis, opinion mining, and learners feedback analysis.

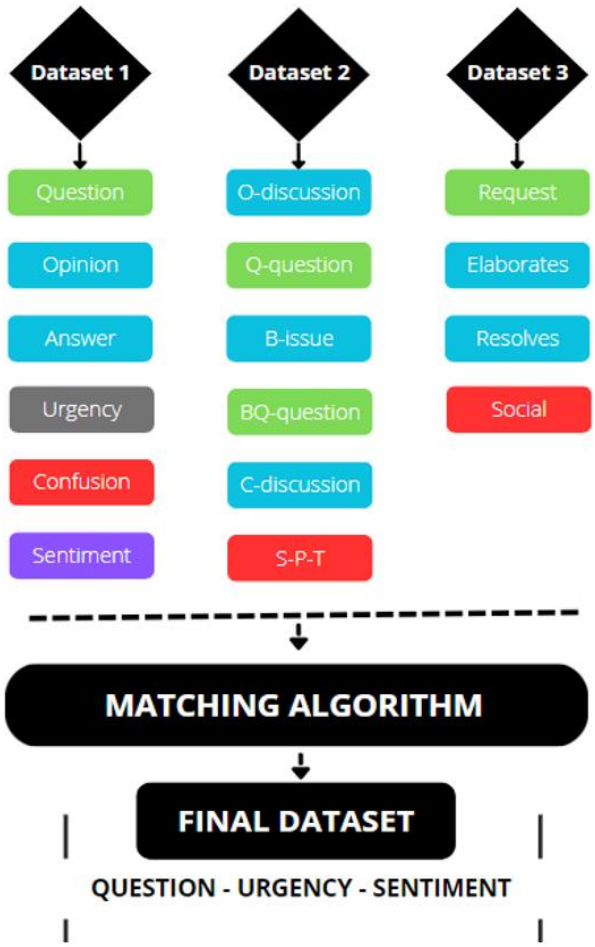

Figure 4. Merging three datasets into one

Each dataset contains a set of classes. To merge them, we used this colour notation to match classes to each other, the green and blue colour refers to the question class, the grey to urgency and purple is the sentiment, the red ones are excluded because we will not make use of them in the classification tasks.

After combining the three datasets, we obtained a dataset composed of 31000 entries with three classes. We performed some data analysis in order to draw a word cloud and investigate the 10 most common words as shown in Figure 5 and Figure 6 respectively.

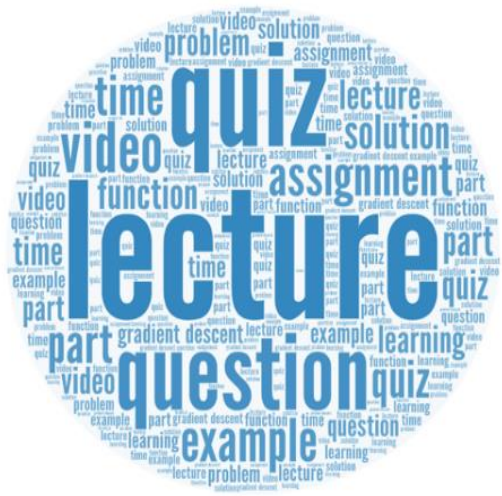

Figure 5. 10 most common used words cloud

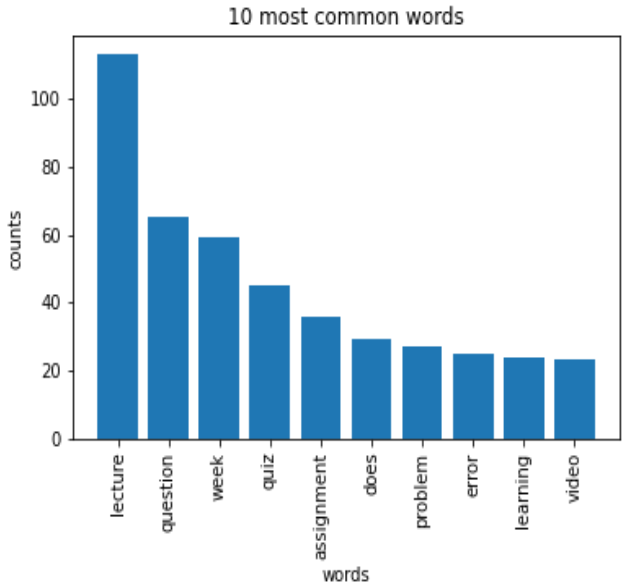

Figure 6. Data frequency

This new dataset, which encompasses the combined and refined information from the three initial datasets was built with the aim of enabling classification tasks specifically optimized for Large Language Models. The comprehensive cleaning and preprocessing steps, coupled with the algorithmic matching of classes, ensured that the final dataset was not only homogenous but also well-suited for classification tasks within the realm of Large Language Models.

This effort to build a new dataset capable of accommodating the requirements of classification tasks with Large Language Models stands as a testament to data quality and the pursuit of effective model training. The resulting dataset serves as a valuable resource, empowering future works in natural language processing and enabling the development of robust and accurate classification models in the field of MOOCS.

4. EXPERIMENTAL SET UP

4.1 Model selection and configuration

We employed state-of-the-art language models for the

classification tasks:

BERT, RoBERTa, GPT-3.5, ELECTRA, and XLNet: These models were selected based on their performance in natural language processing tasks and fine-tuned using transfer learning techniques.

Fine-tuning: Models were fine-tuned on the prepared dataset to adapt them to specific tasks such as urgency classification, topic similarity assessment, and sentiment analysis.

Parameters: Hyperparameters such as learning rate, batch size, and epochs were optimized through preliminary experiments and grid search to maximize model performance.

4.2 Experimental design

4.2.1 Task descriptions

Urgency classification: Models were trained to classify student feedback into urgency levels (e.g., urgent, non-urgent) to prioritize responses and support timely intervention.

Topic similarity: Clustering techniques or similarity measures were employed to group feedback entries based on semantic similarities, aiding in identifying common themes and concerns.

Sentiment analysis: Models were used to analyze the sentiment expressed in student feedback, categorizing sentiments as positive, negative, or neutral to assess overall satisfaction and engagement.

4.2.2 Evaluation metrics

Model performance was evaluated using standard metrics:

Accuracy, Precision, Recall, F1-score: These metrics were computed to measure the effectiveness of models in each classification task.

Cross-validation: Techniques such as k-fold cross-validation ensured robustness and generalizability of results by validating models on different subsets of the dataset.

4.3 Implementation details

Tools: Experiments were conducted using Python programming language and popular libraries such as TensorFlow, PyTorch, and Hugging Face Transformers for model implementations and evaluations.

Hardware: Computations were performed on a GPU-accelerated machine to expedite training and inference tasks.

4.4 Ethical considerations

This study adheres to ethical guidelines regarding data privacy and consent. Measures were taken to anonymize data and protect the identities of individuals participating in forum discussions.

5. FEEDBACK ANALYSIS FRAMEWORK

5.1 Framework design

5.1.1 The urgency framework

The framework suggested groups questions by urgency first then by similarity. Urgency can be portrayed according to the following four levels as shown in Figure 7.

U1 level: Questions that require immediate intervention to prevent lack of knowledge are grouped under this section. This is key in order to be able to follow through the course and could address significant attrition due to lack of memory.

U2 level: High urgency questions encompass matters that deal with understanding issues to avoid confusion in subsequent lessons and are therefore considered crucial to address.

U3 level: Moderate urgency relates to questions that builds on understanding such as applying a certain concept to solve a given problem or analyzing and navigating the complexity of concepts based on what had been seen previously.

U4 level: Low urgency level has to do with questions that test students' ability to evaluate and appreciate a given situation or problem, we decided its urgency is low because such a skill comes later in the learning process and relies on students' maturity and absorption of the whole picture.

5.1.2 The teacher's dashboard

Once the posts are sorted according to urgency levels, they are then grouped by similarity into different categories depending on the topic of the post as shown in Figure 8. Topics may refer to a given lesson or a given chapter in the course.

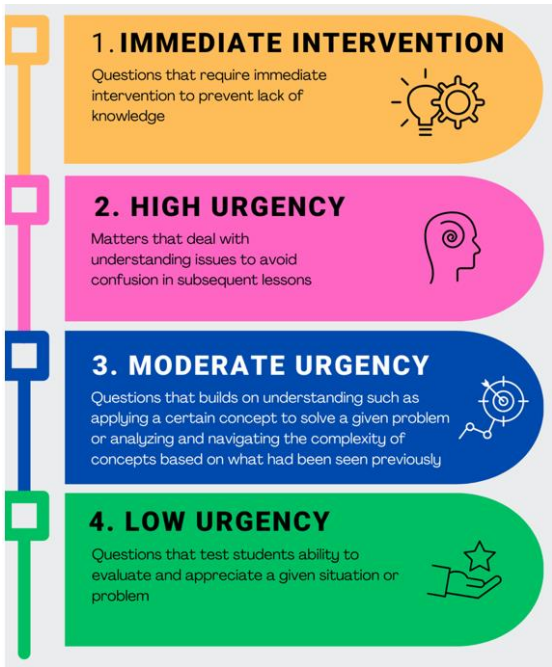

Figure 7. Urgency levels in MOOCs posts

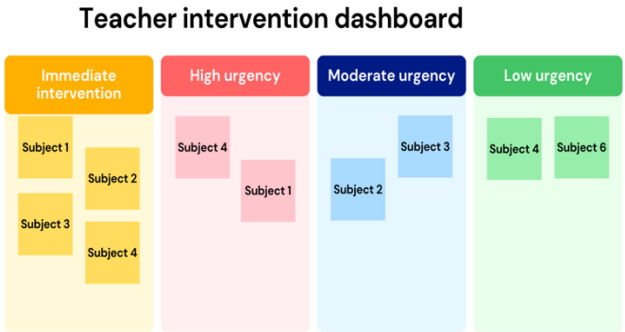

Figure 8. Teacher intervention dashboard grouping forums posts sorted by urgency then grouped by topic similarity

5.1.3 Measuring the classroom temper

The third component of the framework is students' feelings towards the course. In fact, knowing students' satisfaction levels can be a valuable metric to evaluate the classroom

temper. When students express many negative opinions about the course, it could be a red flag about the course and calls for attention to dig more into the roots of such negative opinions.

On the other hand, when students give positive opinions, this implies the course components are satisfactory and little attention should be given to perfecting the course content.

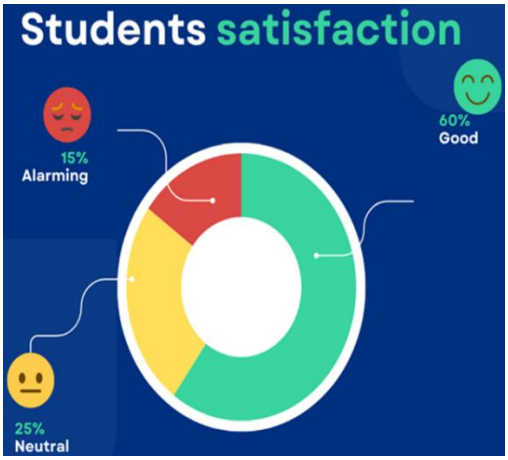

Figure 9. Students' satisfaction dashboard

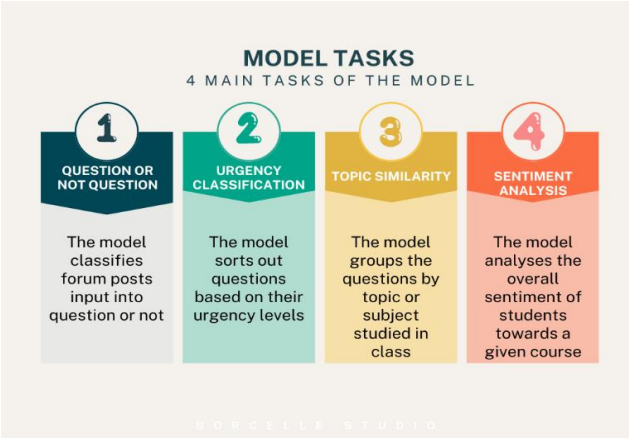

Figure 10. Model's tasks

A simple glimpse at student's satisfaction, as illustrated in the pie chart in Figure 9, would serve as a compass to help teachers and administrators evaluate the content of the course, whether the course fits students' expectations or not, and to decide if significant amelioration needs to be done or not.

5.2 The model's multi tasks

On the basis of previous parts, our model framework is designed to perform the following steps (Figure 10).

•Data is classified into a question or not then the priority level or urgency level of each question is classified.

•Similar questions are grouped together by topics based on their semantic similarity.

•Sentiment analysis is performed to gauge the classroom temper.

The overall process involves a multiclass classification for the first task, clustering to group similar data points, and multi-label classification for sentiment analysis. Therefore, a multi-task learning model was designed to handle three tasks: question-urgency classification, similarity-based clustering, and sentiment analysis. Each task has its own set of labels and categories, and the model is trained to predict the appropriate label for each task given an input data point.

6. IMPLEMENTATION

6.1 Research approach

In this study, we sought to classify feedback following the process described in Figure 11.

Model selection: It consisted of choosing leading linguistic models suitable for multiclass classification tasks, such as BERT, RoBERTa, GPT-3.5, ELECTRA or XLNet. These models have demonstrated robust performance in a variety of natural language processing tasks.

Model architecture: this step involved configuration of the architecture of the models selected for the multiclass classification task. Typically, this implied adding a classification layer on top of the pre-trained language model. It is essential to refine the pre-trained models to suit the specific classification task.

Training: Then, the models were trained on the prepared training dataset. During training, the models were optimized using an appropriate optimizer (namely Adam [19]) and a loss function (in this case categorical cross-entropy [20]). Iterations on training data were carried out for several epochs, adjusting the model's internal parameters to minimize loss.

Hyperparameter tuning: this step consisted of Experimenting with different hyperparametric parameters, such as learning rate, batch size and regularization techniques, to find the optimal configuration that produces the best performance on the validation set.

Evaluation: After training, the models were evaluated on the test set to determine their performance. Also, measures such as accuracy, precision, recall and F1 score were calculated to gauge the models' effectiveness in correctly classifying text samples into the three classes.

Comparison and analysis: the performance of the different models was compared based on the evaluation parameters. Upon analyzing the results, this step helped identify each model's strengths and weaknesses, considering factors such as accuracy, class-specific performance, and overall robustness.

6.2 A multi-task learning model implementation

The model is designed to perform three major tasks as shown in the flowchart in the Figure 12.

Question-urgency: Question classification and priority detection

Figure 11. Model building process

In this section, the aim is to perform a multiclass classification on the final dataset encompassing three classes: "Question", "Urgency" and "Sentiment". Our approach involves training various state-of-the-art language models on this combined dataset. We used Large Language Models because they have shown great ability to grasp complex patterns and semantic representations from textual data. Through the training of these models, we aim to develop accurate and robust classifiers capable of classifying text samples into the three specified classes.

To evaluate and compare the performance of the language models we used in the classification tasks, metrics that are most relevant to our application were used, which are the accuracy, the precision, the recall, and the F1 Score. Higher values for all metrics indicate better classification performance.

The question classification and priority detection task are a multi-class classification problem. To reach our goal, we used our labeled dataset with questions and their corresponding labels indicating whether they are questions or not, and their level of urgency. We used this data to refine a large language model on this dataset, allowing it to learn the patterns and features that distinguish questions from non-questions and then determine the priority level. Table 1 offers a comparison of the different results obtained for each model among the five selected LLMs, pertaining to the metrics mentioned earlier.

Figure 13 shows the accuracy, precision, recall and F1-score of the BERT, RoBERTa, GPT-3.5, ELECTRA and XLNet in classifying student feedback based on urgency levels. All models demonstrated robust performance and achieved high accuracy, precision, recall and F1-score, with RoBERTa peaking to the highest values.

Grouping questions: Clustering questions with similar meaning

To cluster questions with similar meaning, we used semantic similarity score method to build a dataset comprising questions, their similarity scores, and clustering labels. By refining the model on this dataset, it learns to understand and capture the semantic similarity between different questions. We then used the trained model to cluster new questions.

Semantic similarity measures the degree of equivalence in meaning between two text samples. Here is how we used semantic similarity for question grouping:

Unsupervised learning: While supervised learning requires a labeled dataset to train the model, unsupervised learning enables question grouping based solely on the inherent semantic patterns captured by the language model. Lacking these informational labels in our dataset, we opted for unsupervised learning for question grouping by using the pretrained language models to extract sentence embeddings. Then we performed steps 2 and 3 to calculate semantic similarity and put questions into clusters.

Step1: Semantic Similarity Calculation

We calculated the semantic similarity between question pairs using the extracted sentence embeddings. There are several approaches to measure similarity, such as cosine similarity, Euclidean distance, or the use of similarity metrics like Word Mover's Distance (WMD). Word Mover's Distance (WMD) and semantic similarity measures based on pretrained language models are more suitable for capturing semantic similarity between textual sentences [21] for several reasons:

Semantic understanding: Both WMD and pretrained language models have a better understanding of the semantic meaning of words and sentences compared to traditional measures like cosine similarity or Euclidean distance.

Contextual embeddings: Pretrained language models, such as BERT or RoBERTa, use contextual embeddings which provide a more nuanced representation of the semantic content, considering factors like word order, syntactic structure, and semantic relationships.

Sentence-level similarity: Unlike traditional measures that operate at the word level, WMD and semantic similarity measures based on pretrained language capture the overall similarity of sentences by incorporating information from all the words within the sentence, enabling a more comprehensive assessment of semantic similarity.

Handling synonyms and paraphrases: They can recognize that different sentences with varying words can still convey the same meaning. This is particularly beneficial for feedback analysis, where different students may express their thoughts differently but with similar underlying intent.

Table 1. Results comparison of different models on the multiclass [Q-Urgency] classification task

Language Model	Accuracy	Precision	Recall	F1-Score
BERT	0.89	0.90	0.88	0.89
RoBERTa	0.91	0.92	0.91	0.91
GPT-3.5	0.87	0.88	0.86	0.87
ELECTRA	0.90	0.91	0.89	0.90
XLNet	0.88	0.89	0.87	0.88

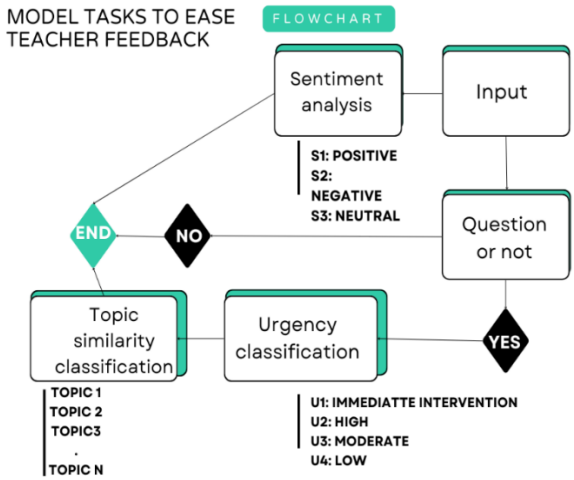

Figure 12. Flow chart of the model's tasks to ease question feedback

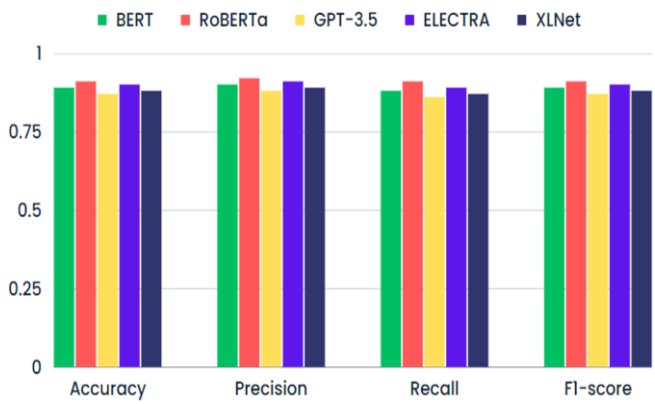

Figure 13. Comparison chart of five LLMs on the multiclass [Q-Urgency] classification task

Step 2: Clustering

We applied a clustering algorithm to the similarity matrix to group questions with similar meaning together. Techniques like hierarchical clustering, k-means clustering, or DBSCAN (Density-Based Spatial Clustering of Applications with Noise) can also be employed. These algorithms group questions based on their similarity scores, forming clusters of questions that share similar meaning.

Both WMD and semantic similarity measures based on pretrained language models can be effective in capturing semantic similarity between textual sentences. However, we found that the computational complexity of WMD is higher compared to the other method, so we opted for the pretrained models.

Table 2 presents the results of Word Mover's Distance (WMD) and semantic similarity measures (BERT) applied to pairs of sentences in our dataset. The "Sentence 1" and "Sentence 2" columns represent the two sentences being compared. The "WMD Score" column displays the Word Mover's Distance score, where lower scores indicate higher semantic similarity. The "BERT Similarity Score" column highlights the similarity score obtained using a pretrained language model like BERT, where higher scores indicate higher semantic similarity.

Sentiment analysis

We finetuned the same LLMs used on the multiclass classification process using our labelled dataset where each feedback is annotated with its corresponding sentiment label.

To capture learners' overall sentiment towards each course, we finetuned the same LLMs used on the multiclass classification process. By training the linguistic models on our data, we sought to exploit the model's inherent language comprehension capabilities to accurately identify and classify the sentiment expressed in the comments (Table 3).

The trained model enabled us to efficiently process and analyze a large volume of feedback data, providing valuable insights into students' satisfaction levels and perceptions of courses.

Figure 14 visualizes the accuracy, precision, recall and F1-score of BERT, RoBERTa, GPT-3.5, ELECTRA and XLNet. All models accurately analyzed sentiment in student feedback and achieved good performance, with RoBERTa peaking at highest values 87%, 88%, 86%, 87% respectively.

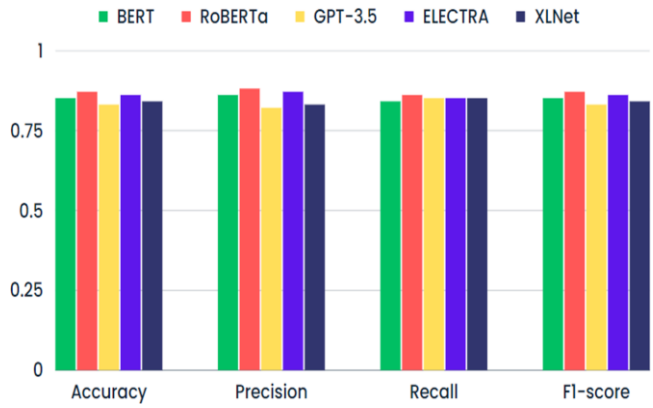

Figure 14. Chart comparison of state-of-the-art LLMs on the sentiment analysis task

6.3 Evaluation and interference analysis of multi-task learning models

In order to get a comprehensive understanding of the strengths and weaknesses of the MTL model, it was primary to select appropriate evaluation metrics. We relied on two evaluation metrics which are:

Task-specific metrics: We calculated task-specific evaluation metrics for each individual task involved in the MTL model.

Task interference analysis: We assessed the potential interference or influence between different tasks in the MTL model. By analyzing how the performance of one task affects the performance of the other tasks, we gained insights into the level of task independence and the overall effectiveness of the learning process.

Table 4 shows the comparison of BERT, RoBERTa, GPT-3.5, ELECTRA and XLNet applied as multitasking learning models. The "Question-Urgency Accuracy" column shows the accuracy of the question and level of priority classification for each model. The "Similarity Performance" column evaluates the questions grouping task. The "Sentiment Analysis accuracy" column represents the accuracy of sentiment analysis. Higher values in all measures indicate better performance.

Table 2. Example of WMD and BERT applied to some sentences of the dataset

Sentence 1	Sentence 2	WMD Score	BERT Similarity Score
"I didn't understand the hierarchical clustering method, can you give an example?"	"Would you please explain the last grouping technique?"	0.25	0.92
"How can we use regular expressions?"	"How to clean data scraped from the web?"	0.42	0.86
" What is an algebraic representation? "	" Need help. Representation theory? "	0.18	0.94
" What impact did the Renaissance have on the intellectual and cultural development of Europe?"	" How did the Renaissance influence the intellectual and cultural advancement of Europe?"	0.35	0.88

Table 3. Results' comparison of state-of-the-art LLMs on the sentiment analysis task

Model	Accuracy	Precision	Recall	F1 Score
BERT	0.85	0.86	0.84	0.85
RoBERTa	0.87	0.88	0.86	0.87
GPT-3.5	0.83	0.82	0.85	0.83
ELECTRA	0.86	0.87	0.85	0.86
XLNet	0.84	0.83	0.85	0.84

Table 4. Results of the three stages of the MTL models

Model	Question-Urgency Accuracy	Similarity Performance	Sentiment Analysis Accuracy
BERT	0.89	0.78	0.85
RoBERTa	0.91	0.82	0.87
GPT-3.5	0.87	0.75	0.83
ELECTRA	0.90	0.80	0.86
XLNet	0.88	0.77	0.84

Table 5. Results of task interference in the MTL models applied to feedback analysis

Model	Task Interference Analysis
BERT	High interference between question-urgency classification and sentiment analysis. Low interference between clustering and other tasks.
RoBERTa	Moderate interference between question-urgency classification and sentiment analysis. Low interference between clustering and other tasks.
GPT-3.5	High interference between question-urgency classification and sentiment analysis. Moderate interference between clustering and other tasks.
ELECTRA	Low interference between all tasks.
XLNet	Moderate interference between question-urgency classification and sentiment analysis. Moderate interference between clustering and other tasks.

Table 5 presents an analysis of task interference in multi-task learning models. The column "Task interference analysis" gives an overview of the level of interference or influence observed between the different tasks in each model. The analysis can be subjective and based on observations of model performance and relationships between tasks. High interference suggests that the performance of one task has a significant impact on the performance of the other tasks, while low interference indicates that the tasks are relatively independent of each other. It is important to note that task interference analysis is a qualitative assessment and may require further research and experimentation to validate the results. The overall performance of these models may vary depending on the specific dataset and the multi-task learning configuration.

7. CONCLUSION

In this study, we used Large Language Models and fine-tuned them on a specific dataset to perform three tasks within discussions forums' classification. Interestingly, we achieved accurate sentiment analysis, question classification, urgency detection, and question grouping. The models learned to capture the semantic nuances and patterns required for each task.

The results of the three tasks could be then integrated into a dedicated dashboard, where teachers and administrators can access real-time information on student pending questions, sentiment trends, track overall learner satisfaction and gain actionable insights to improve course quality and teaching strategies. This comprehensive dashboard could be an invaluable tool for measuring student satisfaction, enhancing the learning experience and enabling teachers and administrators to make data-driven decisions.

The aim of this study is not to present a model that gives the best results, but rather to show a technical architecture that ensures real-time applicability and efficiency with encouraging results for diverse language models.

Chapter 2 The results indicate that using advanced language models for feedback analysis in MOOCs forums can significantly enhance the ability to classify urgency levels, assess topic similarity, and analyze sentiment accurately. These findings underscore the potential of AI-driven approaches to support educators in managing student interactions effectively and improving overall learning experiences in online settings.

As for our perspective for future work, we intend to study the questions in order to get a better understanding of their urgency levels. We strongly believe that urgency levels, if structured carefully, can lead to a better grasp of how teachers should design their courses to smoothen understandability, the urgency framework can serve as a tool to gauge the effectiveness of course material and shed light on areas to improve so as to allow for more students keeping up with the learning pace of MOOCs and ultimately measure how this can contribute directly to lower attrition and drop-out rates in MOOCs settings.

REFERENCES

[1] Marrhich, A., Lafram, I., Berbiche, N., El Alami, J. (2021). Teachers' roles in online environments: How AI based techniques can ease the shift challenges from face-to-face to distance learning. International Journal of Emerging Technologies in Learning, 16(24): 244-254. https://doi.org/10.3991/ijet.v16i24.26367

[2] Marrhich, A., Lafram, I., Berbiche, N., El Alami, J. (2020). A khan framework-based approach to successful MOOCs integration in the academic context. International Journal of Emerging Technologies in Learning, 15(12): 4-19. https://doi.org/10.3991/ijet.v15i12.12929

[3] Gibbs, G. (2010). Using Assessment to Support Student Learning. Leeds Met Press. https://eprints.leedsbeckett.ac.uk/id/eprint/2835/.

[4] Hattie, J., Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1): 81-112. https://doi.org/10.3102/003465430298487

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

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

[7] Amjad, T., Shaheen, Z., Daud, A. (2022). Advanced learning analytics: Aspect based course feedback analysis of MOOC forums to facilitate instructors. IEEE Transactions on Computational Social Systems, 1-9. https://doi.org/10.1109/TCSS.2022.3174640

[8] Nanda, G., Douglas, K.A., Waller, D.R., Merzdorf, H.E., Goldwasser, D. (2021). Analyzing large collections of open-ended feedback from MOOC learners using LDA topic modeling and qualitative analysis. IEEE Transactions on Learning Technologies, 14(2): 146-160. https://doi.org/10.1109/TLT.2021.3064798

[9] Rossi, L.A., Gnawali, O. (2014). Language independent analysis and classification of discussion threads in Coursera MOOC forums. In Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), Redwood City, CA, USA, pp. 654-661. https://doi.org/10.1109/IRI.2014.7051952

[10] Atapattu, T., Falkner, K., Tarmazdi, H. (2016). Topic-wise classification of MOOC discussions: A visual analytics approach. international educational data mining society. In Proceedings of the 9th International Conference on Educational Data Mining, Raleigh, NC, USA, pp. 276-281.

[11] Peng, J.E., Jiang, Y. (2022). Mining opinions on LMOOCs: Sentiment and content analyses of Chinese students' comments in discussion forums. System, 109: 102879. https://doi.org/10.1016/j.system.2022.102879

[12] Wen, M., Yang, D., Rose, C. (2014). Sentiment analysis in MOOC discussion forums: What does it tell us? In Proceedings of the 7th International Conference on Educational Data Mining, pp. 130-137.

[13] Pant, H.V., Lohani, M.C., Pande, J. (2023). Thematic and sentiment analysis of learners' feedback in MOOCs. Journal of Learning for Development, 10(1): 38-54. https://doi.org/10.56059/jl4d.v10i1.740

[14] Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., Galligan, L. (2022). A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. IEEE Access, 10: 56720-56739. https://doi.org/10.1109/ACCESS.2022.3177752

[15] Sun, X., Guo, S., Gao, Y., Zhang, J., Xiao, X., Feng, J. (2019). Identification of urgent posts in MOOC discussion forums using an improved RCNN. In 2019 IEEE World Conference on Engineering Education (EDUNINE), Lima, Peru, pp. 1-5. https://doi.org/10.1109/EDUNINE.2019.8875845

[16] Kumar, V., Troussas, C. (2020). Intelligent Tutoring Systems: 16th International Conference, ITS 2020, Athens, Greece. Springer Nature.

[17] Agrawal, A., Venkatraman, J., Leonard, S., Paepcke, A. (2015). YouEDU: Addressing confusion in MOOC discussion forums by recommending instructional video clips. In National Science Foundation.

[18] Chandrasekaran, M., Ragupathi, K., Kan, M.Y., Tan, B. (2015). Towards feasible instructor intervention in MOOC discussion forums. In Thirty Sixth International Conference on Information Systems, Fort Worth 2015.

[19] Kingma, D.P., Ba, J. (2015). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980

[20] Zhang, Z., Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, pp, 8792-8802.

[21] Kusner, M., Sun, Y., Kolkin, N., Weinberger, K. (2015). From word embeddings to document distances. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Lille, France, pp. 957-966.