Data Science and Engineering (DSE) Record, Volume 6, issue 1.

Tri-Training Based Model Semi-Supervised Aspect-Based Sentiment Analysis: MOOCs Case Study

Kitichart Nukaew

Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand

Kitichart_n@cmu.ac.th

Abstract. Massive Open Online Courses (MOOCs) have seen continuous growth in popularity and rapid expansion. In the instructional design process, receiving feedback from learners is crucial, as it helps tailor the content to better meet learners' needs. The application of NLP models in analyzing learners' feedback is an effective approach for extracting insights from a large volume of comments related to the courses. These models can categorize feedback into three distinct categories: course, instructors, and assessments. Additionally, the models can predict the sentiment of the feedback, determining whether it is positive or negative. In developing these models, semi-supervised learning techniques have been employed to address the challenge of limited data availability. Experimental results indicate that, for feedback categorization, a GRU model combined with tri-training with disagreement yields the highest prediction accuracy. Conversely, for sentiment analysis, a GRU model combined with tri-training produces the best outcomes.

Keywords: text classification, semi-supervised, MOOCs

1 Introduction

Massive open online courses (MOOC) were introduced in 2008 and became one of the popular modes of learning in 2012 because the cost of learning is usually free so anyone with internet can enroll [1]. Compared to traditional learning, MOOCs have more accessibility because of free or very low tuition costs, and MOOCs have more flexibility because they allow to learn at any pace and schedule. MOOCs got high attention, especially after the COVID-19 pandemic which made online learning become the new normal.

Recently, providers of Massive Open Online Courses (MOOCs) have concentrated on offering credentials in fields with well-documented returns on investment, such as data science, computer programming, business, and related disciplines. These credentials are significantly more cost-effective, priced at approximately one-half to one-quarter of the cost of U.S. professional online credentials [2].

In recent years MOOCs have grown rapidly but the dropout rate has been typically very high [3]. Course factors are one of the main groups of factors to make learners drop out of the courses. Course quality and course design are the main reasons for the high dropout rate [4] [5].

The study by Hew et al. [6]. defines MOOCs success as more student satisfaction is more success and revenue from MOOCs. The study proposes a machine learning and sentiment analysis model to predict student satisfaction. Many studies in the past also show that monitoring learner satisfaction in the online course is a key activity for a successful collaborative learning experience [7] [8].

The main problems of MOOCs are low retention and recent enrollment decline. 52% of enrollment students never enter the courseware. From a study in the edx MOOC platform, the retention rate of students in 2017-2018 was only 7%. Although the contents of the MOOCs have a major effect on retention of the course, the instructors' interaction also plays a main part [9].

This study integrates the use of text mining, text embedding, deep learning, and semi-supervised techniques to build and compare the performance of a semi-supervised to develop the category classification models and sentiment analysis models of MOOC learner satisfaction from student feedback from 10 data science-related courses, 10 computer science-related courses, and 10 business-related courses from Coursera which is biggest MOOC provider with the highest number of students. In this research, each sentence from the review will be categorized into 3 aspects: course, instructor, and assessment as broader MOOC-related aspects from previous studies [10], and determine the polarity of each sentence by using the sentiment analysis technique to categorize the sentence into positive and negative sentiment and enhance performance of the model by using the semi-supervised technique.

This study aims to establish a framework for the development of classification models addressing both sentiment analysis and categorical classification within the context of text data from MOOCs related to business, computer science, and data science courses. The proposed framework utilizes a semi-supervised approach, namely tri-training and tri-training with disagreement to reduce the reliance on labeled data while enhancing the model's performance in terms of accuracy, precision, recall, and F1 score.

This study aims to employ various deep learning models, including Artificial Neural Networks (ANN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) networks, and Bidirectional Long Short-Term Memory (BI-LSTM) networks. This framework will incorporate advanced embedding techniques such as Bidirectional Encoder Representations from Transformers (BERT) and leverage semi-supervised technique approaches, specifically tri-training and tri-training with disagreement and to evaluate the impact of applying semi-supervised techniques, including tri-training and tri-training with disagreement, on model performance. This comparison will be conducted relative to models that do not employ semi-supervised techniques. This study aims to employ various deep learning models, including Artificial Neural Networks

(ANN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) networks, and Bidirectional Long Short-Term Memory (BI-LSTM) networks. This framework will incorporate advanced embedding techniques such as Bidirectional Encoder Representations from Transformers (BERT) and leverage semi-supervised technique approaches, specifically tri-training and tri-training with disagreement and to evaluate the impact of applying semi-supervised techniques, including tri-training and tri-training with disagreement, on model performance. This comparison will be conducted relative to models that do not employ semi-supervised techniques.

2 Literature Review

This section explores the development of text classification techniques employed to categorize learners' feedback for courses in Massive Open Online Courses (MOOCs). The initial section provides an introduction to the concept of MOOCs. The following sections review various text classification techniques, including the text embedding model Bidirectional Encoder Representations from Transformers (BERT), the deep learning model for classification such as Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM) networks, Bidirectional Long Short-Term Memory (Bi-LSTM), Gated Recurrent Unit (GRU), the semi-supervised learning methods for model performance enhancement such as tri-training and tri-training with disagreement model.

2.1 Concept of MOOCs

Massive Open Online Courses (MOOCs) were started in 2008 and became one of the popular modes of learning in 2012 because the cost of learning is usually free so anyone with internet can enroll [1]. MOOCs can provide a wide variety of course content, learning methods, and assessments [11]. However, MOOCs have faced the problem of low retention rates and high dropout rates. The study from the edx platform showed the learners from 2012 – 2016 had less than 10% retention rate in 2017 - 2018 [2].

Many researchers try to find the factors that affect retention rates and dropout rates to solve the retention problem. The study conducted by Wang et al. identified several antecedents of dropout rates in Massive Open Online Courses (MOOCs), including psychological, social, personal, and course-related factors, as well as time constraints and unforeseen hidden costs [12]. The study by C.Reparaz et al. [13] performed the logistic regression model to classify MOOCs completers and non-completers. The results indicate that students who completed MOOCs demonstrated a greater ability to self-regulate their learning and exhibited higher levels of perceived effectiveness and engagement with the course content. Researchers have concluded that goal setting, task interest, and academic discipline are the primary predictors of MOOC completion [13]. Another study highlighted that factors significantly influencing student retention in MOOCs include the quality of MOOC content, perceived effectiveness, and the level

of instructor interaction [14]. In recent years MOOCs have grown rapidly but the dropout rate has been typically very high [3]. Course factors are one of the main groups of factors to make learners drop out of the courses. Course quality and course design are the main reasons for the high dropout rate [4] [5].

However, the study from Hew et al. [6] oppose that using dropout rates as a predictor of MOOC success is frequently inaccurate, as many students may not intend to complete the course. The study proposes four key aspects for sentiment analysis to predict student satisfaction: the course instructor, course content, assessment, and schedule. Additionally, learner sentiment encompasses aspects such as course structure, video quality, instructor effectiveness, content relevance, interaction, and assessment methods.

As well as a study by Z.Kastrati et al. [15] identified seven specific factors influencing student opinions about online courses: content, structure, knowledge, skill, experience, assessment, and technology. The study further proposed four broader categories for evaluating MOOC-related aspects: 1) the course, which encompasses both content of the course and structure of the course; 2) the instructor, which includes the instructor's knowledge, skills, and experience; 3) assessment; and 4) technology includes quality of video and voice [15].

In this study, we adopt the broader categorization of MOOCs-related aspects proposed by Z. Kastrati et al. [15] to classify aspects in conjunction with a sentiment analysis model. The aspect categories include course, instructor, assessment, and other. The technology category was excluded due to its insufficient sample size, which rendered it unsuitable for inclusion as a separate class in the categorical aspect model. Consequently, this category was replaced with the "other" class.

Our objective is to utilize broader categorization to establish a framework for investigating learner satisfaction across critical aspects of MOOCs. This framework aims to enhance the quality of MOOCs by focusing on key aspects informed by learner feedback.

2.2 Text Classification

Text classification is the method to predict the class of the given text. Sentiment analysis, also known as opinion mining, is a technique within natural language processing aimed at extracting and analyzing individuals' sentiments, attitudes, and perceptions regarding a particular subject. Sentiment analysis is typically categorized into three levels: aspect level, sentence level, and document level. The aspect level targets the identification of sentiment towards specific entities, the sentence level focuses on sentiments expressed within individual sentences, and the document level encompasses the analysis of sentiments across the entire document [16]. Aspect-based sentiment analysis (ABSA) comprises four primary components: aspect category, aspect term, opinion term, and sentiment polarity. These elements collectively facilitate a detailed examination of sentiments related to specific aspects within a given context [17]. Some

researchers have studied ABSA in compound methods to find only some elements simultaneously such as the study by Liu [18] used BART to create the model for aspect category detection (ACD) and aspect category sentiment analysis (ACSA).

2.2.1 Artificial Neural Network

Artificial Neural Networks (ANNs) represent a fundamental class of neural networks that play an essential role in the domains of machine learning and artificial intelligence. ANNs are structured with multiple layers of nodes arranged in a directed graph, where each layer is fully connected to the subsequent layer.
The architecture of an ANN typically comprises three types of layers: the input layer, hidden layers, and the output layer. The input layer functions as the initial point of entry for data into the network, with each neuron in this layer corresponding to an individual feature of the input data. The hidden layers serve as intermediary processing stages, situated between the input and output layers. The output layer is responsible for generating the final predictions or classifications based on the data transformations performed by the hidden layers.
ANNs are commonly employed as baseline models for evaluating and comparing the performance of other deep learning models.

2.2.2 Long Short-Term Memory

Long Short-Term Memory (LSTM) networks are the one type of recurrent neural network (RNN) which is designed to model sequential data and more effectively capture long-term dependencies compared to traditional RNNs. LSTM networks overcome the limitations of standard RNNs, particularly addressing the issues of vanishing and exploding gradients that impede the learning of long-range dependencies within sequences. As a result, LSTMs have become a foundational architecture in fields such as time-series analysis, and natural language processing, where sequential data is prevalent.
The LSTM model is characterized by its use of memory cells that preserve information across extended sequences. The cell state within an LSTM is regulated by various gates that manage the addition and removal of information. Specifically, the model includes three types of gates: the forget gate, which determines which information from the cell state should be discarded; the input gate, which decides which new information is to be incorporated into the cell state; and the output gate, which governs the output of the current cell and serves as the hidden state for the subsequent time step.
The structure of the LSTM is illustrated in Figure 2.1.

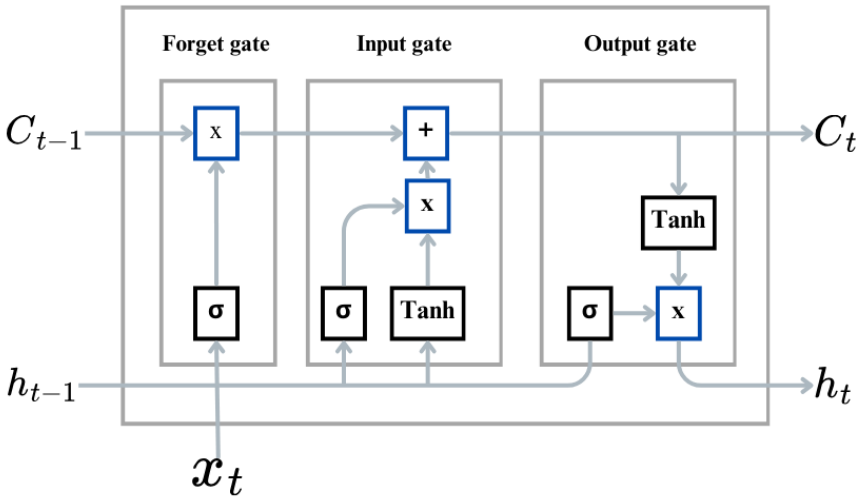

Figure 1 Long Short-Term Memory Architecture

Cell state (C_t) is the memory of the cell, which carries information across different time steps. It is designed to maintain important information over long periods.

Hidden State (h_t) is the output of the LSTM cell at each time step. It is also passed to the next LSTM cell in the sequence.

The input gate (i_t) controls the extent to which new information flows into the cell state. Which is defined by the following equation.

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

Where σ is the sigmoid activation function, W_i is the weight of the matrix, , h_{t-1} is the previous hidden state, $x_t$is the current input, and b_i is the bias term.

Forget Gate (f_t) determines what portion of the cell state from the previous time step should be retained. Which is defined by the following equation.

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

Output Gate (o_t) decides what part of the cell state will be outputted as the hidden state. Which is defined by the following equation.

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

Cell State Update, the cell state is updated based on the input and forget gates, and the new candidate cell state (C_t) is created as follows:

$$C_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t$$

Hidden State Update: the hidden state is updated using the output gate and the updated cell state:

$$h_t = o_t \cdot \tanh(C_t)$$

Recently, the improvement in perspective of the deep learning models has demonstrated significant performance improvements across various domains, including natural language processing (NLP), image processing, and speech recognition. Recurrent neural network (RNN)-based models, such as Long Short-Term Memory (LSTM) networks and Bidirectional LSTM (Bi-LSTM) networks, have proven effective in capturing and modeling sequential information (Minaee, 2021). The Long Short-Term Memory (LSTM) network was introduced as a solution to overcome the vanishing gradient problem inherent in Recurrent Neural Networks (RNNs) (Hochreiter, 1997). LSTM networks show the performance that this model is one of the most effective models for handling natural language processing (NLP) tasks. The research has demonstrated that, in scenarios with limited training data, LSTMs can outperform Bidirectional Encoder Representations from Transformers (BERT) in text classification tasks (Ezen-Can, 2020).

Bidirectional Long Short-Term Memory (Bi-LSTM) is a variant of recurrent neural networks (RNNs) that processes the input sequences in both forward and backward directions. This bidirectional approach enables Bi-LSTM to capture dependencies and information from both directions, enhancing its ability to model complex sequential data (Zhou, 2016, August) (Zhang, 2015, October). While LSTMs are effective at modeling sequential data with long-term dependencies in a single direction, Bi-LSTMs provide a more robust approach by leveraging bidirectional context to capture richer information and dependencies within the sequence.

The Bi-LSTM network consists of two separate LSTM layers: one processes the input sequence in the forward direction and the other processes the input in the backward direction. The outputs of the two layers of Bi-LSTM are then combined, typically through concatenation or addition, to form the final output.
The structure of Bi-LSTM is illustrated in Figure 2.2

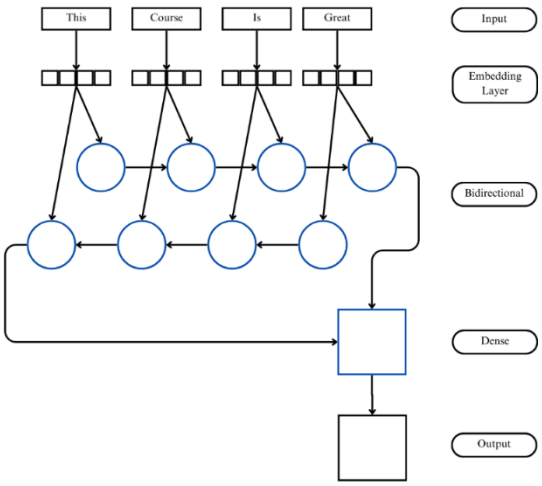

Figure 2 Bidirectional Long Short-Term Memory Architecture

Input Sequence: The input sequence $\{x_1, x_2, \ldots, x_T\}$ is fed into both the forward and backward LSTM layers.
Forward Pass: The forward LSTM processes the input sequence in the forward order, producing the hidden states $\{\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_T}\}$.
Backward Pass: The backward LSTM processes the input sequence in the backward order, producing the hidden states $\{\overleftarrow{h_T}, \overleftarrow{h_{T-1}}, \ldots, \overleftarrow{h_1}\}$.
Combination: At each time step t, the hidden states from both LSTM layers are combined to form the final output:

- Concatenation: $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$
- Addition: $h_t = [\overrightarrow{h_t} + \overleftarrow{h_t}]$

Final Output Sequence: The combined hidden states form the final output sequence $\{h_1, h_2, \ldots, h_T\}$.

2.2.3 Gated Recurrent Unit

Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) architecture that addresses the vanishing gradient problem and aims to provide an efficient mechanism for capturing long-range dependencies in sequential data. GRU was introduced as a simpler alternative to Long Short-Term Memory (LSTM) networks, offering comparable performance with a more streamlined structure.
GRU performs well in sentiment analysis and text classification tasks—the study from Li et al. [26] shown bi-directional GRU model with attention outperformed the other models in sentiment classification with the restaurant review dataset. The other study from Zhang, Xiangsen, et al. [27] used GRU with BERT and multi-head self-attention and outperform the other models in sentiment classification with Yelp and Amazon datasets.

2.3 Semi-supervised training

Semi-supervised learning is interesting in NLP community because unlabeled data have a much higher volume than labeled data [28]. Semi-supervised techniques can help the cost of label problems. Self-training is one of the simplest approaches by training on its own predicted [29].

2.3.1 Tri-Training

The Tri-training model works by leveraging the agreement of three independently trained models to reduce the bias of predictions on unlabeled data. Tri-training employs three classifiers that iteratively label the unlabeled data and retrain each other, enhancing the accuracy and robustness of the model without extensive reliance on labeled datasets. The Tri-training model uses bootstrap sampling and picks predicted data that three independent models predict in the same class and adds these data into the training

dataset to train models again until these three models do not predict data in the same class anymore [30].
Tri-training operates following these steps, let L_1, L_2, L_3 be the 3 different labeled datasets, let U be unlabeled dataset, C_1, C_2, C_3 are classifier models that are trained on L_1, L_2, L_3 dataset, let x is each data in U.

If $C_1(x) = C_2(x)$, add $(x, C_1(x))$ to the training set of C_3.

If $C_2(x) = C_3(x)$, add $(x, C_2(x))$ to the training set of C_1.

If $C_1(x) = C_3(x)$, add $\big(x, C_3(x)\big)$ to the training set of C_2.

Repeat these steps until no data is added to the training set of C_1, C_2, C_3.

3 Data and Methodology

3.1 Research framework

This study presents the research framework for the classification of the 4 factors that affect learner satisfaction, which are course, assessment, instructor, other, and the classification of the polarity of feedback from the learner into 3 sentiments, which are positive, neutral, and negative as depicted. The provided concept framework involves the step of collecting and cleansing data, tokenizing words, and converting them into vectors. The model's accuracy is evaluated by the data in real-world feedback from the learners.

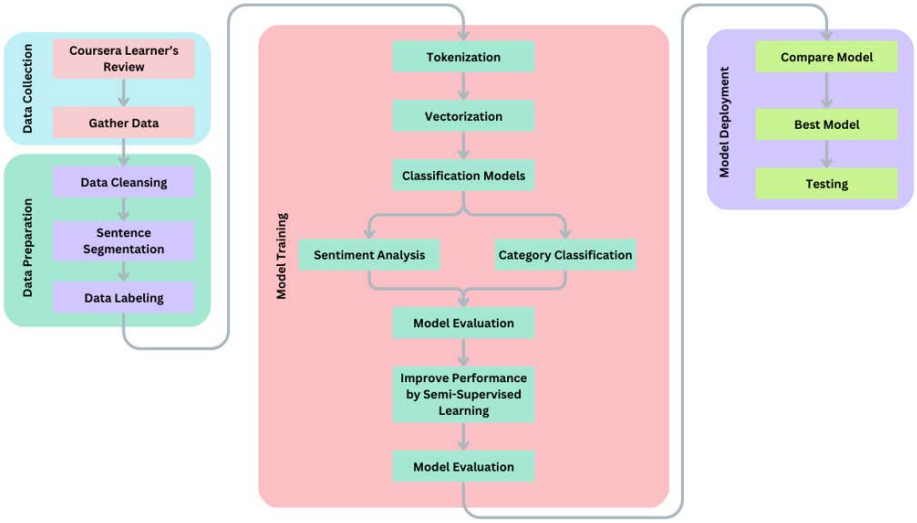

Figure 3 Research Framework

In the data collection and preparation phase, the Beautiful Soup library was employed to extract reviews from the Coursera website. The NLTK library was then utilized to segment the reviews into multiple sentences and preprocess the text by removing punctuation, non-ASCII characters, and converting text to lowercase. Data entries with empty values were excluded, and the dataset was annotated by three linguists to ensure labeling quality.

During the model training phase, the Bidirectional Encoder Representations from Transformers (BERT) framework was used for tokenization and vectorization of the text. Classification models were developed for both sentiment analysis and categorical classification tasks using machine learning and deep learning techniques. To enhance performance, the Synthetic Minority Oversampling Technique (SMOTE) was applied. The models were evaluated based on accuracy, precision, recall, and F1-score to identify candidate models for applying semi-supervised techniques. Semi-supervised techniques, such as Tri-training and Tri-training with Disagreement, were subsequently employed to further improve the models' performance.

In the model evaluation phase, accuracy, precision, recall, and F1-score were used to assess the performance of the models both individually and in combination.

Finally, during the model deployment phase, the models' performances were compared to select the optimal model for both sentiment analysis and categorical aspect classification tasks.

3.2 Data Collection

The dataset utilized in this research was derived from student feedback on courses available on Coursera. Specifically, the data encompassed feedback from 10 data science-related courses, 10 computer science-related courses, and 10 business-related courses, total of 3,000 reviews. Approximately 40% of the dataset was allocated as unlabeled data, while the remaining 60% was designated as labeled data. The labeled dataset was further divided into 60% for the training set and 40% for the evaluation set. To minimize bias, the labeling of the dataset was conducted independently by three linguists. The dataset was composed of the following:

3.2.1 Training Data: A total of 2,777 labeled sentences from course reviews were collected and used for training purposes.

3.2.2 Unlabeled Data for Semi-Supervised Techniques: To enhance model accuracy through semi-supervised learning methods, 2,873 unlabeled sentences from course reviews were included.

3.2.3 Evaluation Data: To assess the model's accuracy, 1,910 labeled sentences from course reviews were utilized.

3.3 Data Preprocessing

Data preprocessing plays a critical role in NLP tasks by cleaning, transforming, and structuring text data to make it suitable for modeling and analysis. For this reason, we used the Natural Language Toolkit (NLTK) to preprocess learners' feedback into a

suitable format for further analysis by deep learning models. The data preprocessing steps can be summarized as follows:

3.3.1 Splitting Sentences: Break up complex sentences into multiple sentences.

3.3.2 Removing punctuation.

3.3.3 Removing non-ASCII Characters.

3.3.4 Lowercase the text.

3.3.5 Removing data that contained empty values.

3.4 Data Filtering

In this step, we aim to classify the sentiment polarity of each sentence in the learner's feedback into 3 classes: positive and negative to reflect the learner's satisfaction with MOOCs. In terms of the category classification of each sentence, we categorized each sentence into 4 classes: course, instructor, assessment, and other [15].

In the process of labeling the polarity of each sentence, three linguists were involved, including one graduate student from the Data Science Consortium. In terms of feedback category, the resulting learner's feedback dataset comprises 2,777 sentences, the sentences categorized into 1,714 course class, 476 assessment class, 452 instructor class, and 135 other class. The bar chart visualizes the distribution of learner's feedback classes in the dataset. In terms of sentence polarity, the sentences were categorized into 1053 negative sentences, 1,610 positive sentences, and 115 neutral sentences.

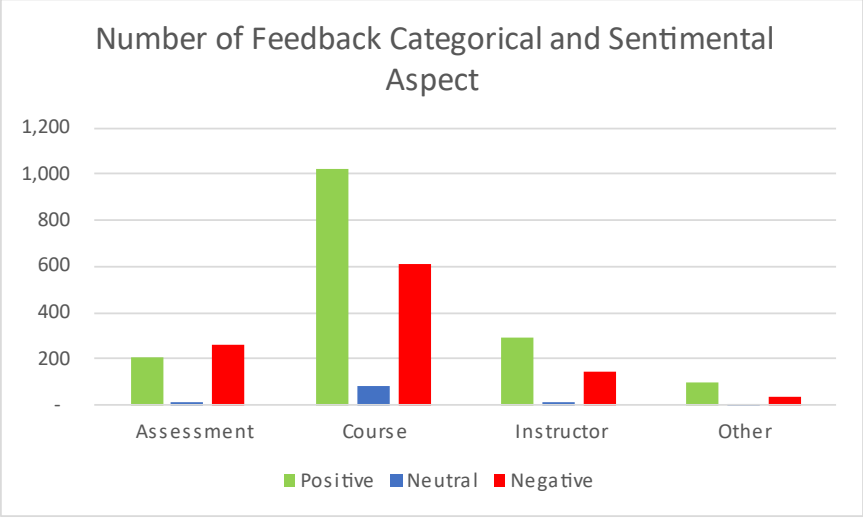

Figure 4 The Number of Feedback Categorical and Sentimental Aspect

3.5 Classification Models

In this phase, we developed machine learning and deep learning models for two distinct tasks: categorical classification and sentiment analysis. For the categorical classification task, the objective was to categorize sentences into three predefined categories: assessment, course, instructor, and other. For the sentiment analysis task, the goal was to categorize the polarity of sentiments into positive, neutral, and negative categories. The machine learning models included Gradient Boosting, Logistic Regression, and Naïve Bayes. The deep learning models employed for these tasks included Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM) networks, and Bidirectional Long Short-Term Memory (Bi-LSTM) networks.

3.6 Imbalance Data Handling

In this phase, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to artificial neural network (ANN) models to evaluate its impact on classification accuracy and determine whether the technique enhances model performance.

3.7 Semi-Supervised Technique

In this phase, we applied semi-supervised techniques named self-training, co-training, tri-training, and tri-training with disagreement to improve the performance of the models in both the categorical aspect classification task and the sentiment analysis task.

3.8 Model Evaluation

Model evaluation involves using multiple metrics to assess the performance of the classification model. In this study, we employed several metrics, including accuracy, precision, F1-score, recall, and confusion matrix to evaluate the effectiveness of the models.

3.8.1 Accuracy

Accuracy is defined as the ratio of correctly predicted samples to the total number of samples. It is calculated by using the following formula:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

3.8.2 Precision

Precision, also known as the positive predictive value, measures the ratio of true positive predictions to the total number of samples classified as positive. It is computed using the formula:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

3.8.3 Recall

Recall, also known as the sensitivity or true positive rate, measures the ratio of true positive predictions to the total number of true positive and false negative. It is computed using the formula:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

3.8.4 F1-Score

The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both aspects. It is defined as:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3.8.5 Confusion Matrix

The Confusion matrix provides a tabular representation of a model's performance by comparing actual and predicted class labels.

4 Evaluation

In this section, we conducted a comparative analysis of the evaluation metrics of various classification models using learners' feedback. Experimental results were evaluated both with and without the application of semi-supervised techniques.

4.1 Comparative Performance of Deep Learning Models and Machine Learning Models

Table 4.1 illustrates that the Artificial Neural Network (ANN) model outperformed the other models in accuracy, precision, and recall. While GRU outperformed the other model in terms of F1-Score in the task of classification of the categorial aspect.

Table 4.1 The performance of the machine learning and deep learning models in the classification task

Model	Accuracy	Precision	Recall	F1-Score
ANN	**82.62%**	**83.00%**	**82.62%**	81.29%
Gradient Boosting	82.15%	80.96%	82.15%	80.83%
GRU	81.62%	82.78%	81.62%	**81.62%**
Logistic Regression	80.68%	81.25%	80.68%	80.76%
Bi-LSTM	80.21%	80.66%	80.21%	80.20%
LSTM	79.53%	81.82%	79.53%	80.14%
Naïve Bayes	57.07%	78.93%	57.07%	61.29%

Table 4.2 demonstrates that the GRU model surpassed other models' accuracy at 86.65%, precision at 87.59%, recall at 86.65, and F1-Score at 86.10% in the sentiment analysis task in the learners' feedback dataset.

Table 4.2 The performance of the deep learning model in the sentiment analysis task

Model	Accuracy	Precision	Recall	F1-Score
GRU	**86.65%**	**87.59%**	**86.65%**	**86.10%**
ANN	85.81%	85.20%	85.81%	85.29%
Logistic Regression	85.65%	85.37%	85.65%	84.86%
Bi-LSTM	85.34%	86.23%	85.34%	85.14%
LSTM	83.93%	83.28%	83.93%	82.63%
Gradient Boosting	83.72%	82.83%	83.72%	82.20%
Naïve Bayes	71.88%	80.32%	71.88%	74.97%

The results of Table 4.1 and Table 4.2 show that deep learning models outperform machine learning models, so we proceed in the oversampling step with deep learning models

4.2 Comparative Performance of Deep Learning After Applying SMOTE

Table 4.3 illustrates that the deep learning models developed without using SMOTE technique have better performance than models developed with SMOTE in the task of classification of the categorial aspect.

Table 4.3 The performance of the deep learning models with SMOTE and without SMOTE in the categorical aspect classification task.

Model	Accuracy	Precision	Recall	F1-Score
ANN	**82.62%**	**83.00%**	**82.62%**	81.29%
GRU	81.62%	82.78%	81.62%	**81.62%**
Bi-LSTM	80.21%	80.66%	80.21%	80.20%
LSTM	79.53%	81.82%	79.53%	80.14%
GRU SMOTE	77.43%	80.38%	77.43%	78.36%
Bi-LSTM SMOTE	76.23%	79.96%	76.23%	77.42%
LSTM SMOTE	72.36%	77.66%	72.36%	73.99%
ANN SMOTE	44.29%	77.02%	44.29%	46.01%

Table 4.4 illustrates that the deep learning models developed without using SMOTE technique have better performance than models developed with SMOTE in the task of classification of the sentiment analysis task.

Table 4.4 The performance of the deep learning models with SMOTE and without SMOTE in the sentiment analysis task.

Model	Accuracy	Precision	Recall	F1-Score
GRU	**86.65%**	**87.59%**	**86.65%**	**86.10%**
Bi-LSTM SMOTE	86.44%	86.48%	86.44%	86.05%
ANN	85.81%	85.20%	85.81%	85.29%
Bi-LSTM	85.34%	86.23%	85.34%	85.14%
LSTM	83.93%	83.28%	83.93%	82.63%
GRU SMOTE	83.35%	85.02%	83.35%	83.75%
LSTM SMOTE	83.30%	86.54%	83.30%	84.50%
ANN SMOTE	83.87%	86.39%	83.87%	84.86%

The results of Table 4.3 and Table 4.4 show that deep learning models without applying SMOTE oversampling technique outperform deep learning models applying SMOTE, so we proceed with deep learning models without using SMOTE.

4.3 Comparative Performance of Deep Learning Applying Different Semi-Supervised Methods

Table 4.5 illustrates that the ANN model developed without using the semi-supervised technique or using the Tri-Training-based semi-supervised technique has better performance than models developed with Self-Training or Co-Training semi-supervised technique in the task of classification of the categorical aspect.

Table 4.5 The performance of the ANN models with different semi-supervised techniques in the categorical aspect classification task

Model	Semi-supervised Technique	Accuracy	Precision	Recall	F1-Score
ANN	-	**82.62%**	**83.00%**	**82.62%**	81.29%
ANN	Co-Training	77.33%	80.16%	77.33%	78.01%
ANN	Self-Training	80.42%	81.98%	80.42%	80.71%
ANN	Tri-Training	82.30%	82.33%	82.30%	**82.18%**
ANN	Tri-Training with Disagreement	78.53%	80.41%	78.53%	79.05%

Table 4.6 illustrates that the ANN model developed without using the semi-supervised technique or using the Tri-Training-based semi-supervised technique has better performance than models developed with Self-Training or Co-Training semi-supervised technique in the task of classification of the sentimental aspect.

Table 4.6 The performance of the ANN models with different semi-supervised techniques in the sentimental aspect classification task.

Model	Semi-supervised Technique	Accuracy	Precision	Recall	F1-Score
ANN	-	85.81%	85.20%	85.81%	85.29%
ANN	Co-Training	86.13%	85.17%	86.13%	84.91%
ANN	Self-Training	84.97%	84.02%	84.97%	83.09%
ANN	Tri-Training	**87.07%**	**86.84%**	**87.07%**	**86.52%**
ANN	Tri-Training with Disagreement	86.81%	86.56%	86.81%	86.21%

The results of Table 4.5 and Table 4.6 show that ANN without applying semi-supervised techniques or ANN with Tri-training-based model has better performance than ANN with self-training or co-training, therefore, we proceed to model development without using self-training or co-training.

4.4 Comparative Performance of Deep Learning After Apply Semi-Supervised Methods

In this part, we use semi-supervised methods such as tri-training and tri-training with disagreement to enhance the performance of models in both classification and sentiment tasks.

Table 8 demonstrates that after applying the semi-supervised technique, the GRU model with the tri-training with disagreement technique surpassed other models, achieving the highest accuracy at 83.04%, precision at 83.72%, recall at 83.04%, and F1-score at 83.06% in the learners' feedback dataset categorical aspect classification task.

Table 4.7 The performance of models after applying the semi-supervised technique in the categorical aspect classification task.

Model	**Semi-supervised Technique**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ANN	-	82.62%	83.00%	82.62%	81.29%
ANN	Tri-Training	82.30%	82.33%	82.30%	82.18%
ANN	Tri-Training with Disagreement	78.53%	80.41%	78.53%	79.05%
Bi-LSTM	-	80.21%	80.66%	80.21%	80.20%
Bi-LSTM	Tri-Training	71.15%	80.06%	71.15%	73.41%
Bi-LSTM	Tri-Training with Disagreement	80.52%	80.54%	80.52%	80.43%
GRU	-	81.62%	82.78%	81.62%	81.62%
GRU	Tri-Training	76.91%	81.03%	76.91%	78.05%
GRU	Tri-Training with Disagreement	**83.04%**	**83.72%**	**83.04%**	**83.06%**
LSTM	-	79.53%	81.82%	79.53%	80.14%
LSTM	Tri-Training	81.47%	82.42%	81.47%	81.59%
LSTM	Tri-Training with Disagreement	82.15%	82.51%	82.15%	81.72%

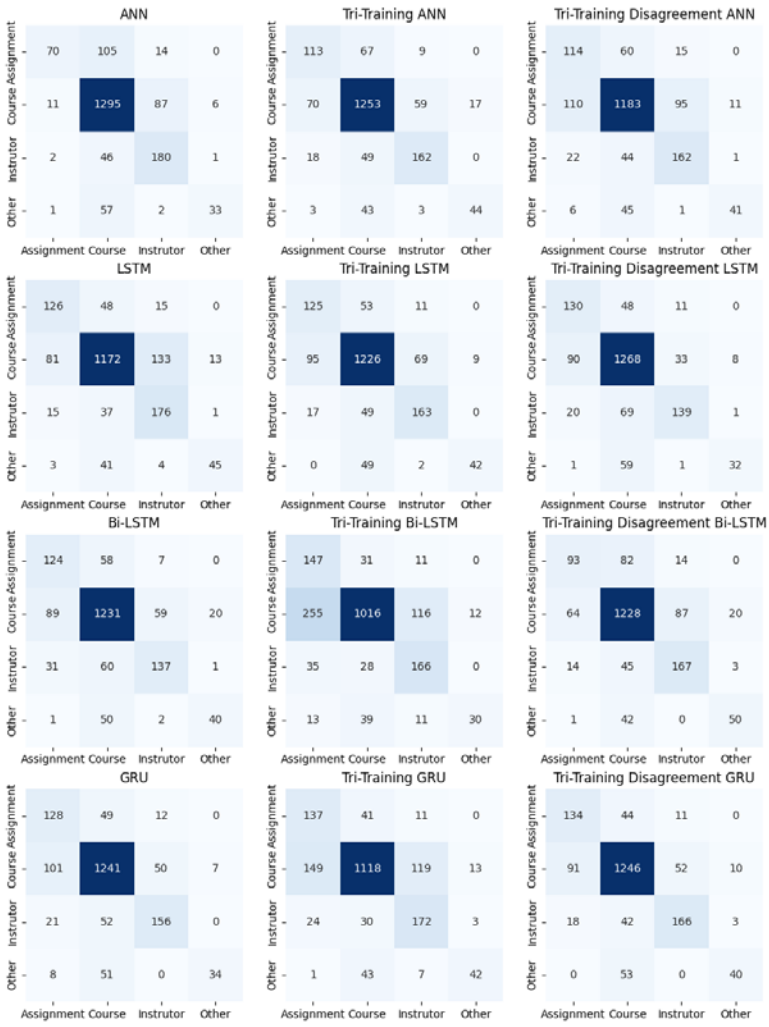

Figure 5 Confusion Matrix of all Categorical Aspect Classification Models

Table 4.8 demonstrates that after applying the semi-supervised technique, the GRU model with the tri-training technique surpassed other models, achieving the highest accuracy, recall, and F1-score while the GRU model without the semi-supervised technique achieved the highest precision in the sentiment analysis task in the learners' feedback dataset.

Table 4.8 The performance of models after applying the semi-supervised technique in the sentiment analysis task

Model	**Semi-supervised Technique**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ANN	-	85.81%	85.20%	85.81%	85.29%
ANN	Tri-Training	87.07%	86.84%	87.07%	86.52%
ANN	Tri-Training with Disagreement	86.81%	86.56%	86.81%	86.21%
Bi-LSTM	-	85.34%	86.23%	85.34%	85.14%
Bi-LSTM	Tri-Training	86.70%	86.64%	86.70%	86.16%
Bi-LSTM	Tri-Training with Disagreement	87.02%	86.92%	87.02%	86.49%
GRU	-	86.65%	**87.59%**	86.65%	86.10%
GRU	Tri-Training	**87.59%**	87.43%	**87.59%**	**87.23%**
GRU	Tri-Training with Disagreement	87.28%	86.95%	87.28%	86.81%
LSTM	-	83.93%	83.28%	83.93%	82.63%
LSTM	Tri-Training	86.18%	85.48%	86.18%	85.68%
LSTM	Tri-Training with Disagreement	85.24%	85.67%	85.24%	84.87%

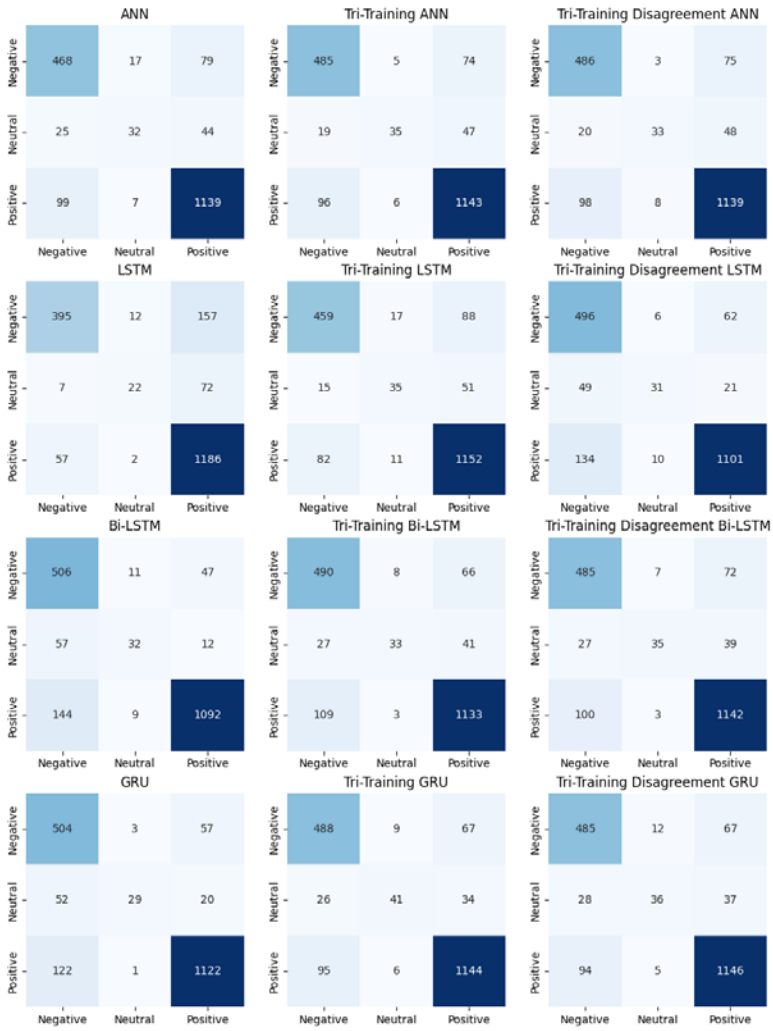

Figure 6 Confusion Matrix of all Sentiment Analysis Models

4.5 Comparative Performance of Deep Learning Model Combinations of Result After Apply Semi-Supervised Methods

Table 4.9 demonstrates the performance of the top 10 model combinations in both sentimental and categorical aspects ranking by accuracy. The results show that the combination of GRU with Tri-training with disagreement for categorical aspect classification and GRU with Tri-training for sentimental aspect classification got the highest accuracy, precision, recall, and F1-score in the task of predicting the sentimental aspect and categorical aspect.

Table 4.9 The performance of combinations of models in both categorical and sentimental aspects

Categorical	Sentimental	Accuracy	Precision	Recall	F1-Score
GRU Tri-Training Disagreement	GRU Tri-Training	**73.14%**	**73.85%**	**73.14%**	**72.95%**
GRU Tri-Training Disagreement	GRU Tri-Training Disagreement	72.72%	73.43%	72.72%	72.47%
GRU Tri-Training Disagreement	ANN Tri-Training Disagreement	72.62%	73.00%	72.62%	72.22%
ANN	GRU Tri-Training	72.62%	73.23%	72.62%	71.21%
GRU Tri-Training Disagreement	Bi-LSTM Tri-Training with Disagreement	72.51%	73.35%	72.51%	72.22%
Gradient Boosting	GRU Tri-Training	72.51%	71.67%	72.51%	71.01%
GRU Tri-Training Disagreement	ANN Tri-Training Disagreement	72.36%	72.75%	72.36%	71.94%
ANN	GRU Tri-Training Disagreement	72.30%	72.86%	72.30%	70.81%
Gradient Boosting	GRU Tri-Training Disagreement	72.30%	71.33%	72.30%	70.73%
GRU Tri-Training Disagreement	Bi-LSTM Tri-Training	72.25%	73.40%	72.25%	72.34%

From the experiments, the structure of the best model can be depicted as follows.

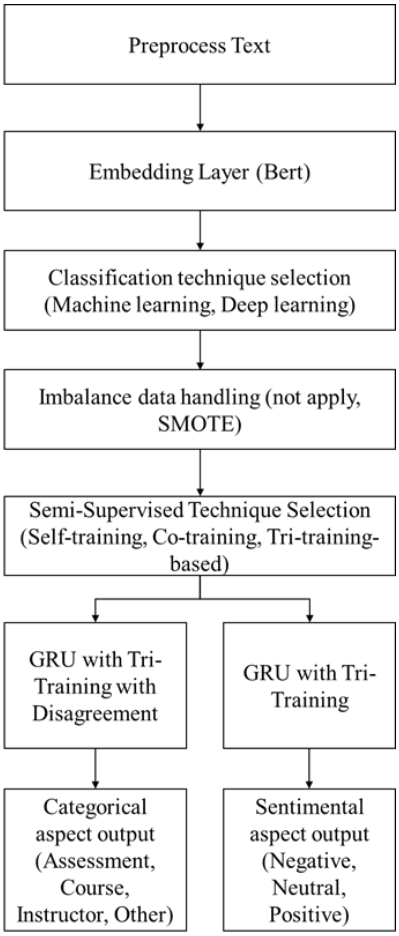

Figure 7 Classification Model for Categorical and Sentimental Aspects

5 CONCLUSION

This study addresses the need for effective analysis of learners' needs in MOOC education platforms by proposing an automatic classification model that includes both categorical classification and sentiment analysis to evaluate learners' feedback. Our research aimed to assess the performance of various deep learning models in classifying feedback within MOOC platforms and to enhance their performance through the application of semi-supervised techniques.

One of the limitations of categorical classification and polarity classification in MOOCs is the classification model requires a huge amount of labeled data to create a good model which costs time and money. The proposed model can solve this issue by

using the semi-supervised approach to get pseudo-label and minimize the cost of labeling.

To achieve this objective, we collected feedback from learners on the Coursera MOOC platform and manually labeled the data for sentiment polarity (Negative, Neutral, Positive) and category (Assessment, Course, Instructor, Other). We then compared the performance of different classification models. The results revealed that the GRU outperformed the others in the sentiment analysis task, with the GRU model achieving the highest accuracy score of 86.65%. In the categorical classification task, the ANN model demonstrated superior performance, reaching an accuracy score of 82.62%.

Additionally, we applied semi-supervised techniques, including tri-training and tri-training with disagreement, to improve model performance. The findings indicate that semi-supervised techniques enhanced the models' performance in both sentiment analysis and categorical classification tasks. Specifically, the GRU model with tri-training achieved the highest accuracy of 87.59% in sentiment analysis, representing a 0.96% improvement over the model without semi-supervised techniques. For the classification task, the Artificial Neural Network (ANN) model with tri-training with disagreement achieved an accuracy of 83.04%, which is a 1.43% improvement compared to the non-semi-supervised models. These results suggest that semi-supervised techniques can be beneficial for this dataset, likely due to its small size and data imbalance.

What about the new finding for MOOC factors?

This study identifies that, among the categorical aspects analyzed, assessment is the only category with a higher proportion of negative sentiment compared to positive sentiment, particularly in technology-intensive courses such as computer science and data science. In contrast, within business-related courses, statements concerning instructors exhibit a higher percentage of positive sentiment compared to other topics. Notably, in data science-related courses, the proportion of negative sentiment regarding the instructor aspect exceeds 60%, whereas in other courses, this percentage remains around 30%.

In conclusion, this research contributes to the field of MOOC education by providing an effective automatic text classification and sentiment analysis model for analyzing learners' needs and feedback. The GRU model with tri-training with disagreement and the GRU model with tri-training demonstrated the best performance compared to other models. Moreover, the research highlights the significant impact of semi-supervised techniques in improving model performance on small and imbalanced datasets.

The limitation of this research in the dataset is limited to 3 groups of subjects (Business, Computer Science, and Data Science) from the Coursera platform, hence the results of the classification performance may be different for classifying the feedback from learners in other subjects and other platforms.

The other limit of this model is the model trained by the data with 1 polarity and 1 category, hence the model is not suitable for multi-label classification.

Future work will focus on employing zero-shot techniques to address the challenges posed by small datasets. This approach aims to leverage open-source models to reduce the costs associated with model training and data labeling.

REFERENCES

[1] L. Pappano, The Year of the MOOC, The New York Times, 2012.

[2] J. &. R.-V. J. A. Reich, "The MOOC pivot.," Science, vol. 363(6423), pp. 130-131, 2019.

[3] S. Halawa, D. Greene and J. Mitchell, "Dropout Prediction in MOOCs using Learner Activity Features," in Proceedings of the second European MOOC stakeholder summit, 37(1), 58-65., 2014.

[4] J. Y. A. H. &. C. D. K. Cheng, "Systematic review of MOOC research in Mainland China," Library Hi Tech, vol. 41(5), pp. 1476-1497, 2022.

[5] Q. Z. W. L. S. a. Z. Y. Jiang, "Empirical research on the specification of design quality in the context of low completion rate of MOOCs," E-Education Research, vol. Vol. 37 No. 01, pp. 51-58, 2016.

[6] K. F. H. X. Q. C. &. T. Y. Hew, "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach," Computers & Education, vol. 145, p. 103724, 2020.

[7] C. N. N. A. C. W. P. L. L. J. R. R. N. &. M. R. M. Gunawardena, "A cross-cultural study of group process and development in online conferences," Distance Education, vol. 22.1, pp. 85-121, 2001.

[8] H. Z. Y. W. J. N. &. W. L. Wu, "Factors associated with occupational stress among Chinese doctors: a cross-sectional survey.," International archives of occupational and environmental health, vol. 83, pp. 155-164, 2010.

[9] K. S. &. E. S. G. R. Hone, "Exploring the factors affecting MOOC retention: A survey study," Computers & Education, vol. 98, pp. 157-168, 2016.

[10] Z. I. A. S. &. K. A. Kastrati, "Weakly supervised framework for aspect-based sentiment analysis on students' reviews of MOOCs," IEEE Access, vol. 8, pp. 106799-106810, 2020.

[11] A. M. F. &. S. T. Yousef, "Reflections on the last decade of MOOC research.," Computer Applications in Engineering Education, vol. 29(4), pp. 648-665., 2021.

[12] W. Z. Y. W. Y. J. &. G. M. Wang, "Factors of dropout from MOOCs: a bibliometric review," Library Hi Tech, vol. 41(2), pp. 432-453, 2023.

[13] C. A.-S. M. &. M. G. Reparaz, "Self-regulation of learning and MOOC retention," Computers in Human Behavior, vol. 111, p. 106423, 2020.

[14] J. &. C. C. Goopio, "The MOOC dropout phenomenon and retention strategies.," Journal of Teaching in Travel & Tourism, vol. 21(2), pp. 177-197, 2021.

[15] Z. K. A. &. H. J. Kastrati, "The effect of a flipped classroom in a SPOC: students' perceptions and attitudes," in Proceedings of the 11th International Conference on Education Technology and Computers, 2019, October.

[16] M. K. M. &. B.-H. A. Birjali, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," Knowledge-Based Systems, vol. 226, p. 107134, 2021.

[17] W. L. X. D. Y. B. L. &. L. W. Zhang, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," IEEE Transactions on Knowledge and Data Engineering, 2022.

[18] J. T. Z. C. L. L. H. &. Z. Y. Liu, "Solving aspect category sentiment analysis as a text generation task," arXiv preprint, vol. arXiv:2110, p. 07310, 2021.

[19] A. S. N. P. N. U. J. J. L. G. A. N. .. &. P. I. Vaswani, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[20] S. &. G.-M. E. C. González-Carvajal, "Comparing BERT against traditional machine learning text classification," arXiv preprint, vol. arXiv:2005.13012., 2020.

[21] S. K. N. C. E. N. N. C. M. &. G. J. Minaee, " Deep learning--based text classification: a comprehensive review.," ACM computing surveys (CSUR), vol. 54(3), pp. 1-40, 2021.

[22] S. &. S. J. Hochreiter, "Long short-term memory," Neural computation, vol. 9(8), pp. 1735-1780, 1997.

[23] A. Ezen-Can, "A Comparison of LSTM and BERT for Small Corpus.," arXiv preprint , vol. arXiv:2009.05451., 2020.

[24] P. S. W. T. J. Q. Z. L. B. H. H. &. X. B. Zhou, "Attention-based bidirectional long short-term memory networks for relation classification," in Proceedings of the 54th annual meeting of the association for computational linguistics, 2016, August.

[25] S. Z. D. H. X. &. Y. M. Zhang, "Bidirectional long short-term memory networks for relation classification," in Proceedings of the 29th Pacific Asia conference on language, information and computation, 2015, October.

[26] L. L. Y. a. Y. Z. Li, "Improving sentiment classification of restaurant reviews with attention-based bi-GRU neural network," Symmetry, vol. 13, no. 8, p. 1517, 2021.

[27] X. e. a. Zhang, "Text sentiment classification based on BERT embedding and sliced multi-head self-attention Bi-GRU," Sensors, vol. 23, no. 3, p. 1481, 2023.

[28] D. C. E. &. J. M. McClosky, "Effective self-training for parsing," in Proceedings of the human language technology conference of the NAACL, main conference, 2006.

[29] Z. H. &. L. M. Zhou, "Tri-training: Exploiting unlabeled data using three classifiers," IEEE Transactions on knowledge and Data Engineering, vol. 17.11, pp. 1529-1541, 2005.

[30] S. &. P. B. Ruder, "Strong baselines for neural semi-supervised learning under domain shift," arXiv, vol. preprint arXiv:1804.09530, 2018.

[31] A. Søgaard, "Simple semi-supervised training of part-of-speech taggers," in Proceedings of the ACL 2010 Conference Short Papers, 2010.

[32] J. S. R. &. M. C. D. Pennington, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, October.

[33] R. &. C. H. Ni, "Sentiment Analysis based on GloVe and LSTM-GRU," in 2020 39th Chinese control conference, 2020.