Which Hammer should I Use? A Systematic Evaluation of Approaches for Classifying Educational Forum Posts

Lele Sha
Centre for Learning Analytics
Faculty of Information Technology
Monash University, Australia
lele.sha1@monash.edu

Mladen Raković
Centre for Learning Analytics
Faculty of Information Technology
Monash University, Australia
mladen.rakovic@monash.edu

Yuheng Li
Faculty of Engineering and Information Technology
University of Melbourne, Australia
yuhengleeeee@gmail.com

Alexander Whitelock-Wainwright
Centre for Learning Analytics
Portfolio of the Deputy Vice-Chancellor Education
Monash University, Australia
alex.wainwright@monash.edu

David Carroll
Centre for Learning Analytics
Portfolio of the Deputy Vice-Chancellor Education
Monash University, Australia
david.carroll@monash.edu

Dragan Gašević
Centre for Learning Analytics
Faculty of Information Technology
Monash University, Australia
dragan.gasevic@monash.edu

Guanliang Chen*
Centre for Learning Analytics
Faculty of Information Technology
Monash University, Australia
guanliang.chen@monash.edu

ABSTRACT

Classifying educational forum posts is a longstanding task in the research of Learning Analytics and Educational Data Mining. Though this task has been tackled by applying both traditional Machine Learning (ML) approaches (e.g., Logistics Regression and Random Forest) and up-to-date Deep Learning (DL) approaches, there lacks a systematic examination of these two types of approaches to portray their performance difference. To better guide researchers and practitioners to select a model that suits their needs the best, this study aimed to systematically compare the effectiveness of these two types of approaches for this specific task. Specifically, we selected a total of six representative models and explored their capabilities by equipping them with either extensive input features that were widely used in previous studies (traditional ML models) or the state-of-the-art pre-trained language model BERT (DL models). Through extensive experiments on two real-world datasets (one is open-sourced), we demonstrated that: (i) DL models uniformly achieved better classification results than traditional ML models and the performance difference ranges from 1.85% to 5.32% with respect to different evaluation metrics; (ii) when applying traditional ML models, different features should be explored and engineered to tackle different classification tasks; (iii) when applying DL models, it tends to be a promising approach to adapt BERT to the specific classification task by fine-tuning its model parameters. We have publicly released our code at `https://github.com/lsha49/LL_EDU_FORUM_CLASSIFIERS`

Keywords

Educational Forum Posts, Text Classification, Deep Neural Network, Pre-trained Language Models

1. INTRODUCTION

In the past two decades, researchers have developed a number of online educational systems to support learning, e.g., Massive Open Online Courses, Moodle, and Google Classroom. Though being widely recognized as a more flexible option compared to campus-based education, these systems are often limited by their asynchronous mode of delivery that may hinder effective interaction between instructors and students and between students themselves [27, 20]. As a remedy, the discussion forum component is often included to support communication between instructors and classmates, so students can create posts for different purposes, e.g., to ask questions, express opinions, or seek technical help. Moreover, in certain cases, instructors rely heavily on the use of a discussion forum to promote peer-to-peer collaboration, e.g., specifying a topic to spur discussions among students.

In this context, the timeliness of an instructor's response to a student post becomes critical. A group of studies has demonstrated that students' learning performance and course ex-

*Corresponding author.

perience were greatly affected by the timeliness of the responses they received from instructors [2, 24, 14]. It is, therefore, critical that instructors monitor the discussion forum to provide timely help to students who need it and ensure the discussion unfolds in a way that benefits all students. However, nowadays, up to tens of thousands of students can enroll in an online course and create a variety of posts that differ by importance, i.e., not all of them warrant instructors' immediate attention. Therefore, it becomes increasingly challenging for instructors to timely identify posts that require an urgent response or to understand how well students collaborate in the discussion space.

To tackle this challenge, various computational approaches have been developed across different courses and domains to classify educational forum posts, e.g., to distinguish between urgent and non-urgent posts [2, 24] or to label posts for different levels of cognitive presence [11, 52]. Typically, these approaches relied upon traditional Machine Learning (ML) models, such as Logistic Regression, Support Vector Machine (SVM), and Random Forest. These models yielded a high level of accuracy, most often due to the extensive efforts that domain experts made to engineer input features. For post classification tasks, such features are linguistic terms describing the post content (e.g., words that represent negative emotions) and the post metadata (e.g., a creation timestamp) [31, 38, 19].

In recent years, Deep Learning (DL) models have emerged as a powerful strand of modeling approaches to tackle data-intensive problems. Compared to traditional ML models, DL models no longer requires the input of expert-engineered features; instead, they are capable of implicitly extracting such features from data with a large number of computational units (i.e., artificial neurons). Particularly, DL models have achieved great success in solving various Natural Language Processing (NLP) problems, e.g., machine translation [48], semantic parsing [22], and named entity recognition [60]. Driven by this, a few studies have been conducted and demonstrated the superiority of DL models over traditional ML models in classifying educational forum posts [24, 10, 59]. For instance, Guo et al. [24] showed that DL models can outperform a decision tree based ML model proposed in [2] by 0.1 (measured by F1 score) in terms of identifying urgent post, while [59] demonstrated that, when determining whether a post contains a question or not, the performance difference between SVM and DL models was up to 0.68 (measured by Accuracy).

Though achieving high performance, DL models have not been justified as an always-more-preferable choice compared to traditional ML models. The reasons are threefold. Firstly, studies investigating the difference in performance between traditional ML and DL models have mostly harnessed a limited set of traditional ML models for comparison, without making extensive feature engineering efforts to empower those traditional ML models. As an example, [59] compared only SVM to a group of DL models, and the SVM model in this study incorporated only one type of features, i.e., the term frequency–inverse document frequency (TF-IDF) score of the words in a post. This implies that the potential of the traditional ML models used in existing studies was not fully explored and the actual performance difference between the two types of models might be smaller than the studies to date have reported on. Secondly, researchers and practitioners often need to deliberately trade off several relevant factors before determining which model they should use in practice, and classification performance is only one of these factors. Other important factors are the availability of human-annotated training data and computing resources [29]. For instance, compared to traditional ML models, DL models demand a much larger amount of human-annotated training data, whose creation can be a time-consuming and costly process. Besides, efficient training of DL models requires access to strong computing resources (e.g., a GPU server), which may be unaffordable to researchers and practitioners with a limited budget. Most traditional ML models, on the other hand, can be easily trained on a laptop. Thirdly, the feature engineering required by traditional ML models plays an important role in contributing to a theoretical understanding of constructs that are not only useful for classification of forum posts, but are also informative about students' discussion behaviors, offering instructors insights on whether their instructional approach works as expected [45, 12, 58].

To assist researchers and educators select relevant models for post classification, this study aims at providing a systematic evaluation of the mainstream ML and DL approaches commonly used to classify educational forum posts. Throughout this evaluation, we advance research in the field by ensuring that: (i) sufficient effort is allocated to design as many meaningful features as possible to empower traditional ML models; (ii) an adequate number of representative ML and DL models is included; (iii) the effectiveness of selected models is examined by using more than one dataset, thus adding to the robustness of our approach to different educational contexts; (iv) all models are compared in the same experimental setting, e.g., with same training/test data splits, and performance reported on widely-used evaluation metrics to provide common ground for model comparison; and (v) the coding schemes used labeling discussion posts are made publicly available to motivate the replication of our study. Formally, the evaluation was guided by the following two **R**esearch **Q**uestions:

- **RQ1** To what extent can traditional ML models accurately classify educational forum posts?
- **RQ2** What is the performance difference between traditional ML models and DL models in classifying educational forum posts?

To answer the RQs, we chose two human-annotated datasets collected at two educational institutions: Stanford University and `Monash University`. We further conducted the evaluation as per the following two classification tasks: (i) whether a post requires an urgent response or not; and (ii) whether the post content is related to knowledge and skills taught in a course. Specifically, to answer RQ1, we first surveyed relevant studies that reported on applying traditional ML models to classify educational forum posts. We hence selected four models that were commonly utilized, i.e., Logistics Regression, Naïve Bays, SVM, and Random Forest. In particular, we collected features frequently employed in the reviewed studies and incorporated them as an input to empower the four traditional ML models in our experiment. Given that these features may play different roles in different classification tasks, we further conducted a feature selection analysis to shed light on the features that must be included in the future application of these models for similar classification tasks.

To answer RQ2, we selected the two widely-adopted DL models, Convolutional Neural Network coupled with Long Short-Term Memory (CNN-LSTM) and Bi-directional LSTM (Bi-LSTM), and compared them to the four selected traditional ML models. Recent studies in DL suggested that the performance of a model adopted for solving an NLP task (CNN-LSTM or Bi-LSTM in our case, denoted as the *task model* for simplicity) can be greatly improved with the aid of state-of-the-art pre-trained language models like BERT [16] in two ways. Firstly, BERT can be used to transform the raw text of a post into a set of semantically accurate vector-based representations (i.e., *word embedding*), which comprise the input information for the task model and enable the model to distinguish among multiple characteristics of a post. Secondly, BERT can adapt itself to capture the unique data characteristics of the task at hand. To this end, BERT couples with the task model and learns the model parameters. In particular, such flexibility has been demonstrated as extremely helpful in the contexts where training data was not sufficient. Therefore, we explored the effectiveness of BERT in empowering the two DL models selected for the experiment. We provide details in Section 3.

Performance of the four traditional ML and two DL models were examined by four evaluation metrics commonly used in classification tasks, i.e., Accuracy, Cohen's κ, Area Under the ROC Curve (AUC), and F1 score. In summary, this study contributed to the literature of the classification of educational forum posts with the following main findings:

- Compared to other traditional ML models, Random Forest is more robust in classifying educational forum posts;
- Both textual and metadata features should be engineered to empower traditional ML models;
- Different features should be designed when applying traditional ML models for different classification tasks;
- DL models tend to outperform traditional ML models and the performance difference ranges from 1.85% to 5.32% with respect to different evaluation metrics;
- Using the pre-trained language model BERT benefits the performance of DL models.

2. RELATED WORK

2.1 Content Analysis of Forum Posts

Across disciplines, educators widely utilize online discussion forums to accomplish different instructional goals. For instance, instructors often provide an online discussion board as a platform for students to ask questions and get answers about course content [12, 57], argue for/against a particular issue and, in that way, engage deeply with course topics [43, 42] or work collaboratively on a course project [49, 13]. In this process, instructors monitor student involvement by reading their posts. At the same time, instructors judge student contributions in the discussion task, e.g., whether students asked a question that relates to course content vs. a question about semester tuition; described their feelings about the discussed problem vs. just rephrased the problem; or clearly communicated their ideas to classmates in a collaborative learning task. Upon identifying posts that do not contribute to the forum at the expected level, the instructor may intervene accordingly. Sometimes, such an intervention needs to be provided immediately (e.g., in a case of a post pointing out the error in the practice exam key).

With the increasing popularity of online discussion forums in the instructional context, educational researchers have become interested in conducting content analysis of students' posts to find evidence and extent of learning processes that instructors aimed to elicit in online discussion. To this end, researchers utilize coding scheme, a predefined protocol that categorizes and describes participants' behaviors representative of the observed educational construct [47, 37], e.g., knowledge building [23, 35], critical thinking [39, 35], argumentative knowledge construction [55], interaction [26, 43], social cues, cognitive/meta-cognitive skills and knowledge, depth of cognitive processing [26, 25], and self-regulated learning in collaborative learning settings [50]. As per the analytical procedure, researchers read student postings and apply a code over a unit of analysis that can be determined physically (e.g., entire post), syntactically (e.g., paragraph, sentence) or semantically (e.g., meaningful unit of text) [15, 47]. Content analysis clearly demonstrated a potential to capture relevant, fine-grained discussion behaviors and provide researchers and educators with warranted inferences made from coding data [46, 47, 28].

Manual content analysis is time-consuming [25], especially in high-enrollment courses with thousands of discussion posts that students create. To automate the process of content analysis and support monitoring of student discussion activity, various computational approaches have been developed for post classification. These approaches relied upon traditional ML models and DL models and handled four common types of post classification tasks: content, confusion, sentiment, and urgency. Below, we expand upon the studies that reported on these tasks.

2.2 Traditional Machine Learning Models

Educational researchers have applied traditional ML models to automate content analysis of online discussion posts for different instructional needs. The ML models we identified in this review are predominantly based on supervised learning paradigm and can be categorized into four general methodological approaches: regression-based (e.g., Logistics Regression [1, 61, 57, 2, 62, 36]), Bayes-based (e.g., Naive Bayes, [5, 4, 36]), kernel-based (e.g., SVM [45, 12, 5, 18, 36, 58, 40, 30]), and tree-based (e.g., Random Forest [5, 2, 36, 31, 19, 38]). These models were designed to predict an outcome variable that represented the meaning of discussion posts across different categories such as confusion, sentiment or urgency. For instance, [12] created an SVM classifier to differentiate between content-related and noncontent-related questions in a discussion thread to help instructors more easily detect content-related discourse across an extensive number of student posts in MOOC, while [1] implemented a Logistic Regression classifier to detect confusion in students' posts and automatically recommend task-relevant learning resources to students who need it. [58] applied SVM to detect student achievement emotions [41] in MOOC forums and studied the effects of those emotions on student course engagement.

In recent years, researchers became increasingly interested in analyzing the expression of urgency (e.g., regarding course content, organization, policy) in a discussion post [2]. For example, [2] developed multiple ML classifiers to identify posts that need prompt attention from course instructors. While researchers mostly implemented supervised ML models, here we also note a small group of studies that reported on using unsupervised methods to classify forum posts, e.g., a lexicographical database of sentiments [36] and minimizing entropy [8].

Traditional ML models built upon textual and non-textual features extracted from students' posts. Textual features characterize content of the discussion post, e.g., presence of domain specific words [61, 44], presence of words reflective of psychological processes [31, 38, 19], term frequency [2, 5], emotional and cognitive tone [12, 57, 40, 58, 7, 34, 1], presence of predefined hashtags [21], text readability index [62], text cohesion metrics [31, 38, 19], and measures of similarity between message text [31, 38, 19]. Non-textual features, on the other hand, include post metadata, e.g., popularity views, votes and responses [12, 45, 36], number of unique social network users [45, 1, 18], timestamp [45, 36], type (post vs. response) [36], variable that signals whether the issue has been resolved or not [61], the relative position of the most similar post [51], variable that signals whether the author of the post is also the initiator of the thread [51], page rank value of the author of current post [51], indicator if a message is the first or last in a thread [38, 31, 19], and structure of the discussion thread [53, 31, 19, 38].

Researchers computed a variety of evaluation metrics to assess performance of these models. Classification accuracy was commonly applied in studies that we reviewed (e.g., [12, 61, 36]). Generally, models achieved classification accuracy of 70% to 90% in classifying forum posts across different levels of content identification, urgency, confusion, and sentiment. We also note that some authors opted for different or additional evaluation metrics, e.g., precision/recall [12, 1, 62], AUC [12, 18], F1 [1, 2], kappa [1, 61, 57]. Across the models, authors utilized a wide range of different validation strategies (e.g., cross validation, train/test split).

We identified two major challenges researchers should be aware of when using traditional machine learning approaches to detect relevant content, confusion, sentiment and/or urgency in a discussion forum. First, traditional machine learning approaches usually involve extensive feature engineering. In the context of post classification, a huge number of textual and non-textual features of a post is practically available to researchers. Features can be generated using different text mining approaches (e.g., dictionary-based, rule-based) and can be even produced using other classifiers (e.g.,[1]). Researchers thus often face a challenge to decide which feature subset to choose to best capture educational problems (e.g., off-topic posting, misinterpreting the discussion task, unproductive interaction with peers) and/or learning process of interest (e.g., knowledge building, critical thinking, argumentation). For this reason, domain and learning experts, including course instructors, learning scientists, and educational psychologists are often needed to define a feature space that aligns with the purpose of an online discussion. Second, works in [12, 57, 4, 2] took a variety of different approaches to validate the classifiers they developed in terms of metrics, datasets, and training parameters which makes it hardly possible to directly compare the performance of these ML models.

2.3 Deep Learning Approaches

To our knowledge, relatively fewer studies attempted to explore the effectiveness of DL approaches in classifying educational forum posts [54, 59, 10, 24, 8, 3, 6]. The DL models adopted by these studies, typically, relied on the use of CNN, LSTM, or a combination of them. For instance, [54] developed a DL model called ConvL, which first used CNN to capture the contextual features that are important to discern the type of a post, and then applied LSTM to further utilize the sequential relationships between these features to assign a label to the post. Through extensive experiments, ConvL was demonstrated to achieve about 81%∼87% Accuracy in classifying discussion posts of different levels of urgency, confusion, and sentiments. In a similar vein, [59] proposed to use Bi-LSTM to better make use of the sequential relationships between different terms contained in a post (i.e., from both of the forward and backward directions). By comparing with SVM and a few DL models, this study showed that Bi-LSTM performed the best in determining whether a post contained a question or not (72%∼75% Accuracy).

It is worth noting that the success of DL models often depends on the availability of a large-amount human-annotated data for model training (typically tens of thousands at least). This, undoubtedly, limits the applicability of DL models in tackling tasks with only a small amount of training data (e.g., a few thousand). Fortunately, with the aid of pre-trained language models like BERT [16], we can still exploit the power of DL models [10]. Pre-trained language models aim to produce semantically meaningful vector-based representations of different words (i.e., word embeddings) by training on a large collection of corpora. For instance, BERT was trained on English Wikipedia articles and Book Corpus, which contain about 2,500 million and 800 million words, respectively. Two distinct benefits were brought by such pre-trained language models: (i) the word embeddings produced by them encode a rich contextual and semantic information of the text and can be well utilized by a task model (e.g., ConvL described above) to distinguish different types of input data; and (ii) a pre-trained language model can be adapted to a specific task by concatenating itself to the task model and further fine-tuning/learning their parameters as a whole with a small amount of training data. For example, [10] showed that BERT was able to boost classification Accuracy up to 83%∼92% when distinguishing posts of different levels of confusion, sentiment, and urgency.

Though gaining some impressive progress, the studies de-

Table 1: The features used as input for traditional ML models. The features used to train models are denoted as *Yes* under the column *Included*

Category	**Feature**	**Description**	**# features**	**Studies used this feature**	**Included**
Textual	# unigrams and bigrams	Only the top 1000 most frequent unigram/bigrams are included.	2000	[12, 57, 40, 2, 56, 62, 51]	Yes
	Post length	# words contained in a post.	1	[58, 45, 36, 62]	
	TF-IDF	The term frequency-inverse document frequency (TF-IDF) of the top 1000 most frequent unigrams.	1000	[2, 5]	
	Automated readability index	A score $\in [0, 100]$ specifying the post readability.	1	[62]	
	LIWC	A set of features denoted as scores $\in [0, 100]$ indicating the characteristics of a post from various textual categories including: *language summary*, *affect*, *function words*, *relativity*, *cognitive process*, *time orientation*, *punctuation*, *personal concerns*, *perceptual process*, *grammar*, *social* and *drives*.	84	[2, 58, 7, 34, 31, 38, 19]	
	Word overlap	The fraction of words that appeared previously in the same post thread.	-	[7]	No
	# domain-specific words	Words selected by expert to characterize a specific subject, e.g., "*equation*" and "*formula*" for Math.	-	[61, 44]	
	LDA-identified words	Words that are specific to topics discovered by applying the topic modeling method Latent Dirichlet allocation.	-	[62, 4, 44]	
	Coh-Metrix	A set of features indicating text coherence (i.e., co-reference, referential, causal, spatial, temporal, and structural cohesion) linguistic complexity, text readability, and lexical category.	-	[31, 38]	
	LSA similarity	A score indicating the average sentence similarity within a message.	-	[31]	
	Hashtags	Hashtags pre-defined by instructors to characterize the type of a post, e.g., #help and #question for confusion detection.	-	[21]	
Metadata	# views	The number of views that a post received.	1	[45, 12, 36, 62, 18]	Yes
	Anonymous post	A binary label to indicate whether a post is anonymous to other students.	1	[45, 1, 18]	
	Creation time	The day and the time when a post was made.	2	[45, 36, 18]	
	# votes	The number of votes that a post received.	1	[45, 12, 36, 62, 18]	
	Post type	A binary label to indicate whether a post is a response to another post.	1	[36, 2]	
	Response time	The amount of time before a post was responded.	-	[18]	No
	# responses	The number of responses that a post received.	-	[45, 12, 36, 18]	
	Discussion status	A binary label to indicate whether the issue has been resolved or not.	-	[61]	
	Comment Depth	A number assigned to a post to indicate its chronological position within a discussion thread.	-	[53]	
	First and Last Post	A binary label to indicate whether the post is the first or the last in a discussion thread respectively.	-	[51, 53]	

scribed above were often limited in providing a systematic comparison between the proposed DL models and existing traditional ML models. In other words, these studies either did not include traditional ML models for comparison [10, 54] or only compared DL models with only one or two traditional ML models and the potential of these traditional ML models might be suppressed due to a limited amount of efforts spent in feature engineering [59]. This necessitates a systematic evaluation of the two strands of approaches so as to better guide researchers and practitioners in selecting models for classifying educational forum posts.

3. METHODS

We open this section by describing the datasets used in our study. Then, we introduce the representative traditional ML models, including the set of features we engineered to empower those models (RQ1), and then describe the two DL models we chose to compare to the four traditional ML models (RQ2).

3.1 Datasets

To ensure a robust comparison between traditional ML and DL models in classifying educational forum posts, we adopted two datasets in the evaluation, briefly describe below.

Stanford-Urgency consists of 29,604 forum posts collected from eleven online courses at Stanford University. These courses mainly cover subjects like medicine, education, humanities, and sciences. To our knowledge, this dataset is one of the few open-sourced datasets for classifying educational forum posts and was widely used in previous studies [57, 2, 5, 10, 56, 24, 21]. In particular, Stanford-Urgency contains three types of human-annotated labels, including the degree of urgency of a post to be handled by an instructor, the degree of confusion expressed by a student in a post, and the sentiment polarity of a post. In line with the increasing research interest in detecting urgent posts [2, 10, 24, 3], we used Stanford-Urgency and focused on determining the levels of urgency of posts in this study. The count of urgent and non-urgent posts is 6,418 (22%) and 23,186 (78%), respectively. Originally, the urgency label was assigned on a Likert scale of $[1, 7]$, with 1 denoting being not urgent at all and 7 denoting being extremely urgent, respectively. Similar to previous studies [2], we pre-processed the data by treating those of value larger than or equal to 4 as urgent posts and those less than 4 as non-urgent posts, and the classification task became a binary classification problem. It is worth pointing out two notable benefits of including Stanford-Urgency: (i) the large number of posts contained in Stanford-Urgency provided sufficient training data for DL models; and (ii) in addition to the text contained in a post, Stanford-Urgency contains rich metadata information about the post, e.g., the creation time of a post, whether the creator of a post was anonymous to other students, the number of up-votes a post received, which enabled us to explore the predictive utility of different types of data.

Moodle-Content was collected by `Monash University`, the dataset contains 3,703 forum posts that students generated in the Learning Management System `Moodle` during their coursework in courses like arts, design, business, economics, computer science, and engineering. The posts were first manually labelled by a junior teaching staff and then independently reviewed (and corrected if necessary) by two additional senior teaching staff to ensure the correctness of the assigned labels. In contrast to Stanford-Urgency, this dataset contains labels to indicate whether a post was related to the knowledge and skills taught in a course or not, e.g., "*What is poly-nominal regression?*" (relevant to course content) vs. "*When is the due date to submit the second assignment?*" (irrelevant). The count of content-relevant and content-irrelevant posts is 2,339 (63%) and 1,364 (37%), respectively. Therefore, similar to the adoption of Stanford-Urgency, we also tackled a binary classification problem here. However, it should be noted that, compared to Stanford-Urgency, the metadata of posts were not available in Moodle-Content.

3.2 Traditional Machine Learning Models

Model Selection. To ensure our evaluation is systematic, we included representative models that emerged in previous studies. As summarized in Section 2.2, the traditional ML models commonly investigated to date can be roughly grouped into four categories, i.e., *regression-based*, *Bayes-based*, *kernel-based*, and *tree-based*. Therefore, we selected one model from each group and explored their capabilities in classifying educational forum posts, namely Logistics Regression, Naïve Bayes, SVM, and Random Forest.

Feature Engineering. Different from previous studies [59, 5], we argued that traditional ML models should involve an extensive set of meaningful features to fully unleash their predictive potential before being compared to DL models, specifically, we expected that ML models demonstrate improved performance when utilising more features. Therefore, we surveyed studies that reported on applying traditional ML models to classify educational forum posts, engineered features following previous studies and incorporated those features into the four traditional ML models, as summarized in Table 1. These features can be classified into two broad categories: (i) *textual* features that are extracted from the raw text of a post with the aid of NLP techniques; and (ii) *metadata* features about a post. As the metadata of posts was not available in Moodle-Content, only textual features were engineered for this dataset, while both textual and metadata features were engineered for Stanford-Urgency. We excluded several types of features from the evaluation, mainly due to the unavailability of the data required to engineer those features, e.g., *# domain-specific words*, and *Hashtags*. As for *LDA-identified words*, *Coh-Metrix*, and *LSA similarity*, we have left these features to be explored in our future work.

Feature Importance Analysis. Previous studies [12, 57, 56] have demonstrated the benefits of feature importance analysis in providing a theoretical understanding of the underlying constructs that are useful to classify educational forum posts, e.g., identifying features that are useful across different classification tasks. Therefore, we adopted the following approach to identify the top k most important features of an ML model:

1. the Chi-squared statistics between engineered features and the target classification labels were computed;
2. each time, the feature of the highest Chi-squared statistic was fed into the model and the feature was kept in the set of input features only if the classification performance had increased;
3. we repeated (2) until k most important features were identified.

3.3 Deep Learning Models

Existing studies on developing DL models to characterize different types of forum posts, typically, involved the use of CNN or LSTM, which motivated us to include the following two DL models to our evaluation:

- **CNN-LSTM** [54, 24, 59]. This model consists of: (i) an input layer, which learns an embedding representation for each word contained in the input test; (ii) a CNN layer, which performs a one-dimensional convolution operation on the embedding representation produced by the input layer and captures the contextual information related to each word; (iii) an LSTM layer, which takes the output of the CNN layer to make use of the sequential information of the words; and (iv) a classification layer, which is fully-connected layer taking the output of the LSTM layer as input to assign a label to the input text.

- **Bi-LSTM** [59, 24, 6]. Though LSTM has been demonstrated as somewhat effective in utilizing the sequential information of long input text, they are limited in only using the previous words to predict the later words in the input text. Therefore, Bi-directional LSTM was proposed, which consists of two LSTM layers, one representing text information in the forward direction and the other in the backward direction to better capture the sequential information between different words. Formally, this model consists of: (i) an input layer (same as CNN-LSTM); (ii) a Bi-LSTM layer; and (iii) a classification layer (same as CNN-LSTM).

Both CNN-LSTM and Bi-LSTM use an input layer to learn the representation of the input text, i.e., embeddings of the words in a post. Instead of learning word embeddings during training, previous studies [17, 33, 32] suggested that pre-trained language models like BERT can be used to initialize embeddings. Such embedding initialization has been demonstrated as an effective way to facilitate a task model to acquire better performance. Therefore, we adopted BERT to initialize the input layer of both CNN-LSTM and Bi-LSTM and, correspondingly, the implemented models are denoted as Emb-CNN-LSTM, and Emb-Bi-LSTM, respectively.

In addition to word embeddings initialization, as suggested in recent studies in the field of NLP [17, 33], we can further couple BERT with a task model (i.e., CNN-LSTM or Bi-LSTM) and adapt BERT to suit the unique characteristics of a task by training BERT and the task model as a whole. In other words, the task model is often concatenated on top of BERT's output for the `[CLS]`, which is a special token used in BERT to encodes the information of the whole input text. The co-training of BERT and the task model enables BERT to fine-tune its parameters to produce task-specific word embeddings for the input text, which further facilitates the task model to determine a suitable label for the input. In fact, this fine-tuning strategy, compared to being used for embedding initialization, has been demonstrated as a more promising approach to make use of BERT. For instance, [10] showed that, even by simply coupling with a classification layer (i.e., the last layer of CNN-LSTM and Bi-LSTM), BERT was capable of accurately classifying 92% forum posts. Most importantly, it should be noted that the parameters of the coupled model can be well fine-tuned/learned with only a few thousand data samples. That means, this fine-tuning strategy enables CNN-LSTM and Bi-LSTM to be also applicable to tasks that deal with only a small amount of data, e.g., Moodle-Content in our case. In summary, we fine-tuned BERT after coupling it with CNN-LSTM (CNN-LSTM-Tuned) and Bi-LSTM (Bi-LSTM-Tuned), respectively. Besides, to gain a clear understanding of the effectiveness of this fine-tuning strategy, we coupled BERT with only a single classification layer (denoted as SCL-Tuned) and compared it with CNN-LSTM-Tuned and Bi-LSTM-Tuned. Table 2 provides a summary of the DL models implemented in this study.

Table 2: The DL models used in this study. Here, SCL denotes Single Classification Layer.

Models	Usage of BERT		Task Model		
	Embedding Initialization	Fine-tuning	CNN-LSTM	Bi-LSTM	SCL
Emb-CNN-LSTM	√		√		
Emb-Bi-LSTM	√			√	
CNN-LSTM-Tuned		√	√		
Bi-LSTM-Tuned		√		√	
SCL-Tuned		√			√

3.4 Experiment Setup

Data pre-processing. Training and testing data were randomly split in the ratio of 8:2. The Python package `NLTK` was applied to perform lower casing and stemming on the raw text of a post after removing the stop words.

Evaluation metrics. In line with previous works in classifying educational forum posts, we adopted the following four metrics, i.e., Accuracy, Cohen's κ, AUC, and F1 score, to examine model performance. We ran each model three times and reported the averaged results.

Model implementation and training. The traditional ML models (i.e., Logistics Regression, Naïve Bays, SVM, and Random Forest) were implemented with the aid of the Python package `scikit-learn` and their parameters were determined by applying grid search and fit the grid to the training data. Note all model hyper-parameters will be documented in the released GitHub repository. The ML models were trained with textual and metadata features for the Stanford-Urgency dataset, and trained with textual features for the Moodle-Content dataset. When applying the method detailed in 3.2 to perform feature importance analysis, we used F1 score as the metric to measure the changed model performance. For both CNN-LSTM and Bi-LSTM, the model parameters are selected to be comparable with similar previous works in [54, 24, 59, 10]. To this purpose, the size of the BERT embeddings used in the input layer was 768 and the number of hidden units used in the final classification layer was 1. We used the activation function sigmoid and L2 regularizer. In CNN-LSTM, the CNN layer was set to have 128 convolution filters with filter width of 5, while the LSTM layer was set to have 128 hidden states and 128 cell states. In Bi-LSTM, the number of the hidden states and cell states in the LSTM cells was both set to 128. For all DL models, (i) 10% of the training data was randomly selected as the validation data; (ii) the batch size was set to 32 and the maximum length of the input text was set to 512; (iii) the optimization algorithm Adam was used; (iv) the learning rate was set by applying the one cycle policy with maximum learning rate of 2e-05; (v) the dropout probability was set to 0.5; and (vi) the maximum number of training epochs was 50 and early stopping mechanisms were used when the model performance on the validation data starts to decrease, and data shuffling was performed at the end of each epoch. The best model is selected based on validation error. For BERT, we used the service provided by `Bert-as-service`[1].

[1] https://github.com/hanxiao/bert-as-service

Table 3: The performance of traditional ML models. The results in bold represent the best performance in each task.

	Stanford-Urgency				**Moodle-Content**			
Methods	**Accuracy**	**Cohen's κ**	**AUC**	**F1**	**Accuracy**	**Cohen's κ**	**AUC**	**F1**
Naïve Bays	0.7536	0.5071	0.7762	0.7844	0.7183	0.4736	0.7210	0.6870
SVM	0.8627	0.7347	0.8630	0.8185	0.7536	0.5900	0.7536	0.7530
Random Forest	**0.8915**	**0.7892**	**0.8916**	**0.8918**	**0.7544**	**0.5927**	**0.7551**	**0.7661**
Logistic Regression	0.8068	0.6287	0.8068	0.7638	0.7339	0.5251	0.7357	0.7547

Table 4: The performance of Random Forest on Stanford-Urgency when using different types of features as input. The results in bold represent the best performance.

Types of Features	**Accuracy**	**Cohen's κ**	**AUC**	**F1**
Textual	0.8639	0.7368	0.8642	0.8652
Metadata	0.8150	0.6442	0.8152	0.8136
Textual + Metadata	**0.8915**	**0.7892**	**0.8916**	**0.8918**

Table 5: The performance of Random Forest when only using the top-10 most important features (Table 6) as input. The fractions within brackets indicate the decreased performance compared to those with all available features as input (Table 3).

Stanford-Urgency				**Moodle-Content**			
Accuracy	**Cohen's κ**	**AUC**	**F1**	**Accuracy**	**Cohen's κ**	**AUC**	**F1**
0.8610 (-3.42%)	0.7315 (-7.31%)	0.8617 (-3.35%)	0.8628 (-3.25%)	0.7175 (-4.89%)	0.5577 (-5.91%)	0.7186 (-4.83%)	0.7358 (-3.96%)

4. RESULTS

Results on RQ1. The performance of the four traditional ML models is presented in Table 3. Across both classification tasks, Random Forest achieved the best performance, as per the calculated evaluation metrics, followed by SVM and Logistics Regression. Naïve Bayes, on the other hand, achieved the lowest performance. Specifically, Random Forest was capable of accurately classifying almost 90% of the forum posts in Stanford-Urgency, and reached an AUC and F1 score of 0.8916 and 0.8918, respectively. Besides, Cohen's κ score achieved by Random Forest for the same dataset was 0.7892, which indicates a substantial (and almost perfect) classification performance. In terms of classifying Moodle-Content, we noticed the overall performance of all models was lower than in Stanford-Urgency. This may be attributed to the lack of metadata features and significantly fewer posts in Moodle-Content than in Stanford-Urgency, making it harder for the models to reveal characteristics of different types of posts in Moodle-Content. Still, Random Forest achieved an overall accuracy, AUC, and F1 score of 0.7544, 0.7551, and 0.7661, respectively, and Cohen's κ score was very close to 0.6, which indicates an almost substantial classification performance.

Before delving into the identification of the most predictive features, we submitted each group of the textual and metadata features to the best-performing ML model (i.e., Random Forest) to depict their overall predictive power. The results are given in Table 4, derived only from Stanford-Urgency due to the unavailability of the metadata features in Moodle-Content. We observe that both textual and metadata features were useful in boosting classification performance, and textual features seem to have had a stronger capacity in distinguishing urgent from non-urgent posts. For instance, when only taking textual features into consideration, the AUC score was 0.8462, which is about 6% higher than that of metadata features (0.8152) and only 5% lower than that when considering both textual features and metadata features.

To gain a deeper understanding of the predictive power of different features, we further applied the method described in Section 3.2 to select the top 10 most important features in both Stanford-Urgency and Moodle-Content, described in Table 6. Here, several interesting observations can be made.

Firstly, almost all of the identified features were textual features, with only one exception observed in Stanford-Urgency, i.e., the metadata feature *# views*. This is in line with the findings we observed in Table 4, i.e., compared to metadata features, textual features tended to make a larger contribution in classifying forum posts. Among those textual features, we should also notice that most of them were extracted with the aid of LIWC. This corroborates with the findings presented in previous studies [31, 38, 19], i.e., LIWC is a useful tool in identifying meaningful features for characterizing educational forum posts.

Secondly, there is little overlap regarding the top ten most important features in the two tasks (only two shared feature, i.e., *LIWC: pronoun* and *LIWC: posemo*). In particular, we note that the number of features was highly related to the context of a classification task. In the Stanford-Urgency case, a number of top features were associated with a sense of stimulation (e.g., *anxiety*, *affect*, *drive*), which represents a subjective representation of urgency. In the Moodle-Content case, features were more associated with a sense of investigation (e.g., *Analytic* and *Understand*). This shows that different classification tasks (i.e., Urgency vs. Content-related) require task-specific features to best capture the task-specific information (i.e., whether the post expressed a sense of ur-

Table 6: The top 10 most important features used in Random Forest. Features shared by the two tasks are in bold.

Stanford-Urgency		**Moodle-Content**	
Features	**Description**	**Features**	**Description**
Metadata: # views	The number of views that a post received.	Post length	# words contained in a post.
LIWC: pronoun	# of the occurrence of all pronouns (e.g., personal and impersonal pronouns)	LIWC: Analytic	A score indicating the formal, logical, and hierarchical thinking patterns in a post
Unigram: they	# of the occurrence of the word "*they*"	LIWC: Tone	A score indicating the emotional tone conveyed in a post
LIWC: number	# of the occurrence of the digital numbers	**LIWC: pronoun**	# of the occurrence of all pronouns (e.g., personal and impersonal pronouns)
LIWC: affect	A score indicating the overal emotion (positive and negative) of a post	LIWC: ppron	# of the occurrence of all personal pronoun (e.g., he, she, me) in a post
LIWC: posemo	A score indicating the positive emotion of a post	Unigram: I	# of the occurrence of the word "*I*"
LIWC: drives	A score indicating the needs, motives, and drives of a post (e.g., references to success and failure)	**LIWC: posemo**	A score indicating the positive emotion of a post
LIWC: power	A score indicating the power of a post (e.g., reference to dominance)	TF-IDF: understand	The TF-IDF score of the word "*understand*"
LIWC: anx	A score indicating the anxiety conveyed in a post	LIWC: affiliation	A score indicating the capacity for enjoying close, harmonious relationships conveyed in a post
LIWC: QMark	# of the occurrence of question mark	LIWC: Exclam	# of the occurrence of exclamation mark

Table 7: The performance of DL models. The results in bold represent the best performance in each task. The fractions within brackets indicate the increased performance compared to the best performance achieved by Random Forest (Table 3).

	Models	**Accuracy**	**Cohen's κ**	**AUC**	**F1**
Stanford-Urgency	1. Emb-CNN-LSTM	0.9203 (3.23%)	0.8192 (3.80%)	0.9201 (3.20%)	0.9203 (3.20%)
	2. Emb-Bi-LSTM	0.9159 (2.73%)	0.8051 (2.01%)	0.9153 (2.66%)	0.9159 (2.71%)
	3. CNN-LSTM-Tuned	**0.9211** (3.32%)	**0.8210** (4.02%)	**0.9221** (3.42%)	**0.9221** (3.40%)
	4. Bi-LSTM-Tuned	0.9210 (3.30%)	0.8196 (3.85%)	0.9208 (3.28%)	0.9210 (3.27%)
	5. SCL-Tuned	0.9210 (3.31%)	0.8206 (3.98%)	0.9215 (3.35%)	0.9219 (3.38%)
Moodle-Content	6. CNN-LSTM-Tuned	**0.7934** (5.17%)	**0.6230** (5.11%)	**0.7952** (5.32%)	**0.7993** (4.33%)
	7. Bi-LSTM-Tuned	0.7854 (4.11%)	0.6220 (4.93%)	0.7901 (4.64%)	0.7913 (3.29%)
	8. SCL-Tuned	0.7716 (2.29%)	0.6092 (2.77%)	0.7733 (2.42%)	0.7803 (1.85%)

gency).

Moreover, when solely using the top 10 features as an input, the performance of Random Forest was 3.25%∼7.31% lower than the performance obtained after incorporating all available features (Table 5). This finding hence confirms that while the traditional ML models can achieve good classification performance using only the top 10 best features, there is still potential for improvement when using more features. Hence, researchers should attempt to apply more features to fully unleash traditional ML models' capability.

Results on RQ2. The performance of the implemented DL models is presented in Table 7. As Moodle-Content contained only 3,703 labeled posts, that was likely to be insufficient to support the training of CNN-LSTM or Bi-LSTM from scratch. Therefore, we only implemented the fine-tuned models, i.e., CNN-LSTM-Tuned, Bi-LSTM-Tuned, and SCL-Tuned on Moodle-Content. Several observation can be derived based on the results in Table 7.

Firstly and unsurprisingly, DL models uniformly achieved a better performance than traditional ML models. This corroborates findings reported in [59, 54, 33, 24]. DL models are, therefore, superior to traditional ML models in terms of capturing the characteristics of a dataset and obtaining better classification results. However, we should note that the performance difference between traditional ML models and DL models was not that large. Specifically, the best-performing model CNN-LSTM-Tuned achieved an improvement of only 3.32% in Accuracy, 4.02% in Cohen's κ, 3.42% in AUC, and 3.40% in F1 score. In particular, the Cohen's κ score was 0.8210, which suggests an almost perfect classification performance.

Secondly, contrasting findings reported in [59], we found that CNN-LSTM slightly outperform Bi-LSTM in most cases (i.e., Row 1 vs. Row 2, Row 3 vs. Row 4, and Row 6 vs. Row 7 in Table 7). Thirdly, instead of using BERT for embedding initialization, the classification model would achieve better performance by fine-tuning BERT by coupling it with the task model and training the coupled model as a whole (i.e., Row 1-2 vs. Row 3-4 in Table 7), though the improvement was rather limited, e.g., less than 1% when comparing to that of Emb-CNN-LSTM and CNN-LSTM-Tuned on Stanford-Urgency. Fourthly, we showed that in Stanford-Urgency, by simply coupling BERT with a single classification layer (SCL-Tuned, Row 5 in Table 7), the classification performance was almost as good as those derived by coupling BERT with more complex DL models like CNN-LSTM and Bi-LSTM (Row 3-4 in Table 7). This implies that, BERT can capture the rich semantic information hidden behind a post, which can be used to deliver adequate classification performance even by employing a single classification layer.

5. DISCUSSION AND CONCLUSION

The classification of educational forum posts has been a longstanding task in the research of Learning Analytics and Educational Data Mining. Though quite some previous studies have been conducted to explore the applicability and effectiveness of traditional ML models and DL models in solving this task, a systematic comparison between these two types of approaches has not been conducted to date. Therefore, this study set out to provide such an evaluation with aiming at paving the road to researchers and practitioners to select appropriate predictive models when tackling this task. Specifically, we compared the performance of four representative traditional ML models (i.e., Logistics Regression, Naïve Bays, SVM, and Random Forest) and two commonly-applied DL models (i.e., CNN-LSTM and Bi-LSTM) on two datasets. We further elaborate on several implications that our work may have on the development of classifiers for educational forum posts. We also list limitations to be addressed in future studies.

Implications. Firstly, the performance difference between traditional ML models and DL models was not as large as reported by previous studies (e.g., [59]). More specifically, we showed that traditional ML models were often inferior to DL models in terms of only 1.85% to 5.32% decrease in classification performance measured by Accuracy, Cohen's κ, AUC, and F1 score. This finding implies that, when researchers and practitioners have no access to strong computing resources and, for this reason, cannot utilize DL models, they can still achieve acceptable classification performance by using traditional ML models, as long as those ML models incorporate carefully-crafted features.

Secondly, our results demonstrate that the performance of Random Forest classifier is more robust compared to other traditional ML models. This implies that other more advanced tree-based ML models (e.g., Gradient Tree Boosting [9]) might be worth exploring to achieve even higher classification performance. Besides, given that the most important feature in Stanford-Urgency was *# views* (Table 6) and the models' performance in Moodle-Content might be suppressed due to the unavailability of metadata features, it may be worth paying special attention to acquiring and using metadata features when applying traditional ML models. Another finding suggests that little overlap was detected between the top 10 most important features selected in each of the two classification tasks (Table 6). This implies when tackling a classification task, features should be designed to suit the unique characteristics of the task and fit the theoretical model utilized to annotate data (e.g., with predefined coding scheme). This aligns with findings presented in [31, 38, 19], in different phases of cognitive presence, different importance scores were obtained for the same features. Lastly, researchers and practitioners may wish to take advantage of pre-trained language models like BERT when developing DL models. Our experiment showed that BERT can be effectively used in two ways, i.e., (i) to initialize the word embeddings of the post text as the input for a task model; or (ii) to suit the needs of the specific classification task by coupling itself with the task model and then fine-tuning model parameters. Particularly, the second way enables DL models to be applicable to tasks that deal with only a small amount of human-annotated data, like in Moodle-Content).

Limitations. Firstly, the evaluation presented in this study focused only two classification tasks, i.e., Stanford-Urgency and Moodle-Content. To further increase the reliability of the presented findings, more tasks should be included and investigated, e.g., determining the level of confusion that a student expressed in a forum post or whether the sentiment contained in the post is positive or negative [10, 54]. Secondly, a few types of features were not included when exploring the capabilities of traditional ML models in our evaluation, e.g., *# domain-specific words* and *LDA-identified words*. To accurately depict the upper bound of the performance of traditional ML models in classifying educational forum posts, it would be worthy to recruit domain experts to further engineer and make use of these features. Thirdly, we should notice that the DL models used in our evaluation (i.e., CNN-LSTM and Bi-LSTM) only utilized the raw text of a post as input and left the metadata features untapped. Given that metadata features have been demonstrated of great importance in the application of traditional ML models, future research efforts should also be allocated to design more advanced DL models that are capable of using both the raw text of a post and the metadata of the post for classification.

Lastly, we acknowledge that, due to the scope of this study, we did not attempt to investigate the reasons causing the performance difference between traditional ML models and DL models, e.g., whether the two categories of models misclassified the same types of messages. In the future, we will further investigate whether the performance difference between traditional ML models and DL models can be attributed to their model structures and explore potential methods to boost their classification performance, e.g., collecting additional forum posts to continue the pre-training of BERT before coupling it with a downstream classification model.

6. REFERENCES

[1] A. Agrawal, J. Venkatraman, S. Leonard, and A. Paepcke. Youedu: addressing confusion in mooc discussion forums by recommending instructional video clips. 2015.

[2] O. Almatrafi, A. Johri, and H. Rangwala. Needle in a haystack: Identifying learner posts that require urgent response in mooc discussion forums. *Computers & Education*, 118:1–9, 2018.

[3] L. Alrajhi, K. Alharbi, and A. I. Cristea. A multidimensional deep learner model of urgent instructor intervention need in mooc forum posts. In *Intelligent Tutoring Systems*, pages 226–236. Springer International Publishing, 2020.

[4] T. Atapattu, K. Falkner, and H. Tarmazdi. Topic-wise classification of mooc discussions: A visual analytics approach. *EDM*, 2016.

[5] A. Bakharia. Towards cross-domain mooc forum post

classification. In *Learning@Scale*, pages 253–256, 2016.

[6] F. Brahman, N. Varghese, and S. Bhat. Effective Forum Curation via Multi-task Learning. page 8, 2020.

[7] A. Caines, S. Pastrana, A. Hutchings, and P. J. Buttery. Automatically identifying the function and intent of posts in underground forums. *Crime Science*, 7(1):19, 2018.

[8] J. Chen, J. Feng, X. Sun, and Y. Liu. Co-training semi-supervised deep learning for sentiment classification of mooc forum posts. *Symmetry*, 12(1):8, 2020.

[9] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794, 2016.

[10] B. Clavié and K. Gal. Edubert: Pretrained deep language models for learning analytics. *arXiv preprint arXiv:1912.00690*, 2019.

[11] A. Cohen, U. Shimony, R. Nachmias, and T. Soffer. Active learners' characterization in mooc forums and their generated knowledge. *British Journal of Educational Technology*, 50(1):177–198, 2019.

[12] Y. Cui and A. F. Wise. Identifying content-related threads in mooc discussion forums. In *Learning@Scale*, pages 299–303, 2015.

[13] D. D. Curtis and M. J. Lawson. Exploring collaborative online learning. *Journal of Asynchronous learning networks*, 5(1):21–34, 2001.

[14] M. Dascalu, S. Trausan-Matu, D. S. McNamara, and P. Dessus. Readerbench: Automated evaluation of collaboration based on cohesion and dialogism. *International journal of computer-supported collaborative learning*, 10(4):395–423, 2015.

[15] B. De Wever, T. Schellens, M. Valcke, and H. Van Keer. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & education*, 46(1):6–28, 2006.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] J. Dong, F. He, Y. Guo, and H. Zhang. A commodity review sentiment analysis based on bert-cnn model. In *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, pages 143–147. IEEE, 2020.

[18] L. Feng, G. Liu, S. Luo, and S. Liu. A transferable framework: Classification and visualization of mooc discussion threads. In *International Conference on Neural Information Processing*, pages 377–384. Springer, 2017.

[19] M. Ferreira, V. Rolim, R. F. Mello, R. D. Lins, G. Chen, and D. Gašević. Towards automatic content analysis of social presence in transcripts of online discussions. In *LAK*, pages 141–150, 2020.

[20] E. L. Fu, J. van Aalst, and C. K. Chan. Toward a classification of discourse patterns in asynchronous online discussions. *International Journal of Computer-Supported Collaborative Learning*, 11(4):441–478, 2016.

[21] S. A. Geller, N. Hoernle, K. Gal, A. Segal, A. X. Zhang, D. Karger, M. T. Facciotti, and M. Igo. # confused and beyond: detecting confusion in course forums using students' hashtags. In *LAK*, pages 589–594, 2020.

[22] E. Grefenstette, P. Blunsom, N. De Freitas, and K. M. Hermann. A deep architecture for semantic parsing. *arXiv preprint arXiv:1404.7296*, 2014.

[23] C. N. Gunawardena, C. A. Lowe, and T. Anderson. Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of educational computing research*, 17(4):397–431, 1997.

[24] S. X. Guo, X. Sun, S. X. Wang, Y. Gao, and J. Feng. Attention-based character-word hybrid neural networks with semantic and structural information for identifying of urgent posts in mooc discussion forums. *IEEE Access*, 7:120522–120532, 2019.

[25] N. Hara, C. J. Bonk, and C. Angeli. Content analysis of online discussion in an applied educational psychology course. *Instructional science*, 28(2):115–152, 2000.

[26] F. Henri. Computer conferencing and content analysis. In *Collaborative learning through computer conferencing*, pages 117–136. Springer, 1992.

[27] K. F. Hew and W. S. Cheung. Students' and instructors' use of massive open online courses (moocs): Motivations and challenges. *Educational research review*, 12:45–58, 2014.

[28] D. H. Jonassen and H. Kwon. Communication patterns in computer mediated versus face-to-face group problem solving. *Educational technology research and development*, 49(1):35, 2001.

[29] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[30] A. Khan, I. Ibrahim, M. I. Uddin, M. Zubair, S. Ahmad, A. Firdausi, M. Dzulqarnain, and M. Zaindin. Machine learning approach for answer detection in discussion forums: An application of big data analytics. *Scientific Programming*, 2020, 2020.

[31] V. Kovanović, S. Joksimović, Z. Waters, D. Gašević, K. Kitto, M. Hatala, and G. Siemens. Towards automated content analysis of discussion transcripts: A cognitive presence case. In *LAK*, pages 15–24, 2016.

[32] X. Li, L. Bing, W. Zhang, and W. Lam. Exploiting bert for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*, 2019.

[33] X. Li, H. Zhang, Y. Ouyang, X. Zhang, and W. Rong. A shallow bert-cnn model for sentiment analysis on moocs comments. In *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, pages 1–6. IEEE, 2019.

[34] M. Lui and T. Baldwin. Classifying user forum participants: Separating the gurus from the hacks, and other tales of the internet. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 49–57, 2010.

[35] R. M. Marra, J. L. Moore, and A. K. Klimczak. Content analysis of online discussion forums: A comparative analysis of protocols. *Educational Technology Research and Development*, 52(2):23, 2004.

[36] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, I. Estévez-Ayres, and C. D. Kloos.

Sentiment analysis in moocs: A case study. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 1489–1496. IEEE, 2018.

[37] J. Mu, K. Stegmann, E. Mayfield, C. Rosé, and F. Fischer. The acodea framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International journal of computer-supported collaborative learning*, 7(2):285–305, 2012.

[38] V. Neto, V. Rolim, R. Ferreira, V. Kovanović, D. Gašević, R. D. Lins, and R. Lins. Automated analysis of cognitive presence in online discussions written in portuguese. In *Proceedings of the 13th European Conference on Technology Enhanced Learning*, pages 245–261. Springer, 2018.

[39] D. R. Newman, B. Webb, and C. Cochrane. A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology*, 3(2):56–77, 1995.

[40] A. Ntourmas, N. Avouris, S. Daskalaki, and Y. Dimitriadis. Comparative study of two different mooc forums posts classifiers: analysis and generalizability issues. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–8. IEEE, 2019.

[41] R. Pekrun. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, 18(4):315–341, 2006.

[42] R. Rabbany, S. Elatia, M. Takaffoli, and O. R. Zaïane. Collaborative learning of students in online discussion forums: A social network analysis perspective. In *EDM*, pages 441–466. Springer, 2014.

[43] M. Raković, Z. Marzouk, A. Liaqat, P. H. Winne, and J. C. Nesbit. Fine grained analysis of students' online discussion posts. *Computers & Education*, 157:103982, 2020.

[44] A. Ramesh, S. H. Kumar, J. Foulds, and L. Getoor. Weakly supervised models of aspect-sentiment for online course discussion forums. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 74–83, 2015.

[45] L. A. Rossi and O. Gnawali. 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)*, pages 654–661. IEEE, 2014.

[46] L. Rourke and T. Anderson. Exploring social communication in computer conferencing. *Journal of Interactive Learning Research*, 13(3):259–275, 2002.

[47] L. Rourke and T. Anderson. Validity in quantitative content analysis. *Educational technology research and development*, 52(1):5, 2004.

[48] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, and S. Jain. Machine translation using deep learning: An overview. In *2017 international conference on computer, communications and electronics (comptelix)*, pages 162–167. IEEE, 2017.

[49] H.-J. So. When groups decide to use asynchronous online discussions: collaborative learning and social presence under a voluntary participation structure. *Journal of Computer Assisted Learning*, 25(2):143–160, 2009.

[50] M. Sobocinski, J. Malmberg, and S. Järvelä. Exploring temporal sequences of regulatory phases and associated interactions in low-and high-challenge collaborative learning sessions. *Metacognition and Learning*, 12(2):275–294, 2017.

[51] C. Sun, S.-w. Li, and L. Lin. Thread structure prediction for mooc discussion forum. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*, pages 92–101. Springer, 2016.

[52] X. Wang, D. Yang, M. Wen, K. Koedinger, and C. P. Rosé. Investigating how student's cognitive behavior in mooc discussion forums affect learning gains. *EDM*, 2015.

[53] Z. Waters, V. Kovanović, K. Kitto, and D. Gašević. Structure matters: Adoption of structured classification approach in the context of cognitive presence classification. In *Proceedings of the 11th Asia Information Retrieval Societies Conference*, pages 227–238. Springer, 2015.

[54] X. Wei, H. Lin, L. Yang, and Y. Yu. A Convolution-LSTM-Based Deep Neural Network for Cross-Domain MOOC Forum Post Classification. *Information*, 8(3):92, Sept. 2017. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[55] A. Weinberger and F. Fischer. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & education*, 46(1):71–95, 2006.

[56] A. F. Wise, Y. Cui, W. Jin, and J. Vytasek. Mining for gold: Identifying content-related mooc discussion threads across domains through linguistic modeling. *The Internet and Higher Education*, 32:11–28, 2017.

[57] A. F. Wise, Y. Cui, and J. Vytasek. Bringing order to chaos in mooc discussion forums with content-related thread identification. In *LAK*, pages 188–197, 2016.

[58] W. Xing, H. Tang, and B. Pei. Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of moocs. *The Internet and Higher Education*, 43:100690, 2019.

[59] Y. Xu and C. F. Lynch. What do you want? applying deep learning models to detect question topics in mooc forum posts? In *Wood-stock'18: ACM Symposium on Neural Gaze Detection*, pages 1–6, 2018.

[60] V. Yadav and S. Bethard. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*, 2019.

[61] D. Yang, M. Wen, I. Howley, R. Kraut, and C. Rose. Exploring the effect of confusion in discussion forums of massive open online courses. In *Learning@Scale*, pages 121–130, 2015.

[62] Z. Zeng, S. Chaturvedi, and S. Bhat. Learner affect through the looking glass: Characterization and detection of confusion in online courses. *EDM*, 2017.