What do you want? Applying deep learning models to detect question topics in MOOC forum posts?

Yiqiao Xu
yxu35@ncsu.edu
North Carolina State University Computer Science Department
Raleigh, NC

Collin F. Lynch
cflynch@ncsu.edu
North Carolina State University Computer Science Department
Raleigh, NC

ABSTRACT

Students in Massive Open Online Course (MOOC) usually share opinions, express concern, and seeking help by participating in discussion via online forums. However, it's impossible for instructor stuff to go through all the posts in details, find all seeking help posts in time, and response based upon their content accurately, due to the large volume of registrants. In this study, we proposed an identification framework based on a combination of convolutional neural network and long short term memory model (CNN-LSTM), and Bi-directional LSTM (BiLSTM) to automatically classify whether a post seeking help, and identify what kind of question it asking according to the content. In addition, according to the fact that a large proportion of tokens in our MOOC corpus not included in the pre-trained word embedding model. We compared the word embedding weight pre-trained by Wikipedia(GloVe) and the MOOC corpus. This study suggested that our model can potentially significantly increase the efficiency of monitoring MOOC students discussion in real-time.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

MOOC, neural networks, LSTM, text classification, discussion forum

ACM Reference Format:
Yiqiao Xu and Collin F. Lynch. 2018. What do you want? Applying deep learning models to detect question topics in MOOC forum posts?. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Massive open online courses (MOOCs) have gain become increasingly popular over the last decade and have delivered new learning opportunities worldwide. Unlike traditional classroom environments, almost all of the relevant classroom communication takes place on one public channel, the online forum. For most students the forum is the only channel that they have to seek instructor support, ask questions, or connect with their classmates. In face-to-face classes, by contrast, students can seek help in person, can email or post questions online, engage in informal discussions with known peers, or just meet to copy the answers. As a consequence MOOC forums tend to be not only larger but far more active than online forums in traditional courses and they offer a rich and mostly-complete record of the students' learning processes, confusions, social interactions, and concerns. The large volume and high activity of MOOC forums, however can make them relatively unwieldy, making it difficult for instructors to efficiently triage posts to find meaningful questions among the chatter, and to provide appropriate answers, or to determine when student-provided answers are inappropriate. As a result enterprising students may go without receiving the support that they need and be discouraged from continuing the course. Our goal in this work is to address this issue by developing automated deep-learning models to classify posts by type. The research questions are:

- RQ1: Does deep learning model help question triage in MOOC discussion forum?
- RQ2: What kind of deep learning model perform well on MOOC question triage task?
- RQ3: Whether the popular well known word embedding model(GloVe) perform well for MOOC corpus?

To answer RQ1 and RQ2, we proposed three deep learning models to classified MOOC students forum posts and compared with our previous work based upon Support Vector Machine. We first considered question identification as a binary classification problem, and then, according to the text content, classify questions into technique question, course logistic question, and course content question. In order to examine different neural network structure, we applied LSTM as the baseline model, added a CNN layer before LSTM to increase the efficiency of the model and capture the adjacent features in text, and used a Bi-directional LSTM to capture both forward and backward sequential information. As for the third research question, we applied the above three models to classify questions with GloVe word embedding and training a word2vec embedding model with the MOOC corpus itself. The main contribution in this work are: first, we applied deep learning model to capture students seeking help in the discussion forum more accurate; second, we provided a deeper insight of the massive forum for instructors by categorizing question into three types; third, as we found during experiment, there are many tokens missing in

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

pre-trained word embedding model, we compared the results of different word embedding.

2 BACKGROUND

Many researchers have applied NLP techniques to provide a deep insight of students discussion forum and identify the relationship between student success and their activities on the discussion forums. Wen et al[18], for example examined the content of forum posts in MOOCs that include students' attitude towards the course and whether they completed it. They found that student posts with positive motivation words and personal pronouns have lower probability to drop out this course. They also reported a significant correlation between the number of daily drop out students with sentiment features. Wise and Cui [5, 19, 20] categorized MOOC discussion forum posts into whether related to course material content at thread level based upon logistic regression. Then they built content-related and non-related forum social network according to the reply relations. Finally, they analyzed the correlation between course-related posts and students' final grade. They found that both course-related posts and non-related posts were positive relevant to students' final grade. Wang et al [15, 16] modeled students' learning behaviours by classified their MOOC forum posts into four classes based upon ICAP framework [3] and investigated how different cognitive behaviours influenced their learning gains. They found that students who exhibited more high-order thinking behaviours learned more and had deeper participants on the forum.

To categorize forum posts, Lin et al. [11] built a cascade Support Vector Machine model which classified six categories of posts combined with heavy feature engineering. Other than tf-idf as linguistic features, they also extracted sentence position, POS tags, post length, and parent posts category, etc as their structural features. In their results, the highest classification f-score was the announcement class 0.717 and lowest was conflict class 0.132.

Moreover, recent years, deep learning models has achieved state-of-art results in many Natural Language Processing tasks. Deep learning attempts to learn high-level features from data in an increment manner. Long short term memory neural network(LSTM)[7] is a specific type of recurrent neural network(RNN)[4] that designed for modeling long range sequence dependency. LSTM model has been an crucial role in many sequence learning problems, especially text classification. Many models achieved a good performance based on LSTM[12, 21–23]. For example, Zhou et al. [22] combined convolutional neural network(CNN) and LSTM to predict the sentiment polarity of movie reviews. A few researches applied deep learning models to solve pedagogy problems. Wei et al.[17] developed a cross-domain forum post classification with Stanford public MOOCs dataset to solve the cold start problem for the beginning of each course offering. They applied CNN-LSTM model to achieve three independent binary classification task according to posts' confusion/urgency/sentiment. Their model contained four layers: word embedding layer, convolution layer, LSTM layer and the output layer. Finally, they received accuracy as 81.45, 85.91, and 86.69 for the three binary classification tasks: confusion/urgency/sentiment.

Different from them, we added an additional max-pool layer after convolution layer in order to select the most important features. We also proposed a Bi-directional LSTM model to obtain information both from forward and backward. From the results, we found that the Bi-LSTM performed better than CNN-LSTM.

3 DATA

In this study, we used ”Big Data in Education” MOOC provided by The Teachers College at Columbia University and hosted on the Coursera (BDE 2013) and EdX (BDE 2015) platforms in 2013 and 2015 respectively. BDE is offered as an 8 week course that includes material from a graduate-level course on educational data mining and the analysis of big data in education. This curriculum introduces students to basic data collection and data analysis methods such as visualization and clustering. The students learn how and when to do educational data mining and learning analytic on data. The course was structured around weekly lecture videos and individual quizzes. In the 2013 class, 778 students made at least one post or comment on the discussion forum producing a total of 603 discussion threads consisting of 4259 posts in total. In 2015, 519 students produced 625 discussion threads with a total of 2056 posts. We first manually annotated all of the posts and comments separating them into question and non-question. Then among all question posts, we annotated question posts into: technique question, course logistic question, and course content question. Table 2 shows the distribution of items in each group.

Category	BDE2013	BDE2015
Question	972	360
Non-Question	3287	1696

Table 1: Number of questions of each course

Category	BDE2013	BDE2015
Course Content Question (CQ)	666	133
Response of CQ (C)	1292	228
Technique Question (TQ)	201	161
Response of TQ (T)	406	362
Course Logistic Question(LQ)	105	66
Response of LQ (L)	205	100

Table 2: Summary of each class

4 METHODS

To answer the research questions, we developed a process to automatically organize the discussion forum by finding students seeking help actions. To answer *RQ1: Does deep learning model help question triage in MOOC discussion forum?* and *RQ2: What kind of deep learning model perform well on MOOC question triage task?*, first, we annotated datasets to identify questions and the three types of questions. Then we built classifiers with deep learning models based upon CNN-LSTM and Bi-directional LSTM model. Finally, as for *Whether the popular well known word embedding model(GloVe) perform well for MOOC corpus?*, we compared the model performance with the two different word embedding methods.

4.1 Annotation

We conducted an annotation process where we categorized posts into questions and non-questions and further categorized the posts by topic into questions or replies about course content, about techniques or technical support, or other relevant issues (e.g. scheduling). This annotation was conducted by three experienced researchers who annotated all of the posts in our two datasets. Two researchers annotated all the posts in the BDE 2013 dataset, and two annotated the 2015 dataset. Then lead author annotated all posts in both datasets. Both annotation processes followed the same sequence with the annotators marking up a sample set of training posts, discussing disagreements, and repeating until a basic level of agreement was reached. They then annotated the remainder of the posts independently. We then calculated the final inter-rater agreement using Cohen's kappa (κ) [2]. We achieved a κ of 0.81 for BDE2013 and 0.71 for BDE2015, which indicated a very good agreement between two individuals [14].

4.2 Word Embedding

During text prepossessing, we removed punctuation except the question mark, hyperlinks and non ascii code characters. We set the length to all input as the same($max - length$) by cutting off longer sentences and fulfilling 0 for shorter sentences in the training set, based on the fact that the convolutional layer requires fixed length of input.

The first word embedding model is pre-trained by Pennington et al. [13] based upon 6 billions tokens of Wikipedia 2014 and Gigaword 5. Then we trained the MOOC word embedding with this 'BDE' MOOC corpus based on skip-gram[6], which takes every word in a large corpus(focus word), and defines a window to choose surrounding words to feed into a neural network. After training, the neural network is able to predict the probability of each to appear around the focus word.

4.3 CNN-LSTM Model

Figure 1 describes the architecture of CNN-LSTM neural network, which consists of two important component: convolutional neural network and long short term memory neural network.

Convolutional Layer. Convolutional neural network is a deep feed-forward fully connected artificial neural networks and use a variation of multilayer perceptions designed to require minimal prepossessing[9]. The vector v_i represent d-dimension word vector for the i-th word in a L length sentence, and m represent a filter for the convolution operation. Thus, the window vector w within k length for each position j in the sentence becomes:

$$w_j = [v_j, v_{j+1}, ..., v_{j-k+1}]$$

To generate a feature map c, a filter x does convolution operation with the window vectors at each position:

$$c_j = f(w_j.x + b)$$

Where b is a bias term shared by all units in the same layer, and f is a non-linear transformation function.

The one dimensional convolutional layer worked as a slides window to filter vector over a sequence and detecting adjacent features. If instructors trying to answer the questions between 'Could you help me process the data in Python?' and 'Could you help me process the data in R?', the programming language is a significant feature to capture by CNN.

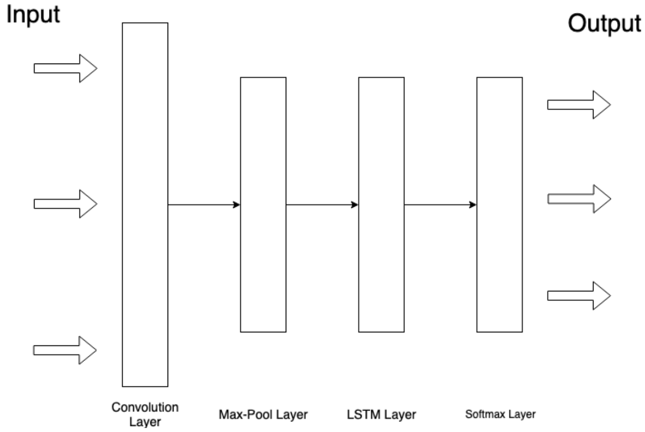

Figure 1: The architecture of CNN-LSTM for sentence modeling

A max-over-pooling layer is used to select the most or the top k-most significant features after the convolution operation.

Long-Short-Term-Memory Neural Network. Recurrent Neural Network(RNN) is able to propagate sequential information based on a chain based network architecture. However, it's impossible for standard RNN to learn the long term dependencies because of the large gap between two timesteps. Long short term memory networks are a special type of Recurrent Neural Network(RNN), which capable of learning long-term dependencies[7]. This model is designed to avoid the long-term dependency problem by remembering the information for long period time. LSTM architecture has a series of repeat standard RNNs as a unit for each timestep, which consists of input gate(h_l), forget gate(f_l), output date(o_l), and sigm and tanh are applied element-wise. The following equation is LSTM in this study[8],

$$h_t^{l-1}, h_{t-1}^l, c_{t-1}^l \rightarrow h_t^l, c_t^l$$

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{2n,4n} \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$

$$c_t^l = f \odot c_{t-1}^l + i \odot g$$

$$h_t^l = o \odot \tanh(c_t^l)$$

However, CNN and LSTM are individually limited to their model abilities. I.e., CNN captures important features that are well depicted through the convolution operation but it will lose the order input information. The advantage of LSTM is insensitivity to the long term period, which allows LSTM to 'remember' the previous information. So it can address the limitation of CNN by sequential modeling text input across sentence with order.

Category	Definition	Example
Course Content Question (CQ)	Questions related to course topics or homework completion.	How does the teacher want the answer to be? percentages or decimal?
Technique Question (TQ)	Technical questions about coding or software issues.	I am running OS X 10.7.5. Do I have to upgrade for this? I have Java 7-Update 45, isn’t that enough? I have the same proble as Paul.
Course Logistic Question(LQ)	Questions about course logistics.	Since no one seems to be jumping in here, I’ll start by asking how many weeks of lectures have already been recorded?

Table 3: Defination and Example of each question type

4.4 Bi-directional LSTM

LSTM captures the sequential information, however, unidirectional LSTM will only use the previous words to predict following words. But, bidirectional LSTM have both information from previous state and the future state. Figure 2 [1] shows the structure of bidirectional LSTM. It has two layers of LSTM, one access information in forward direction and the other one access in the backward direction. So, these networks capture both past and future context. Figure 3 [10] shows our Bi-LSTM model for classification. The input layer and the output layer are the same with CNN-LSTM model.

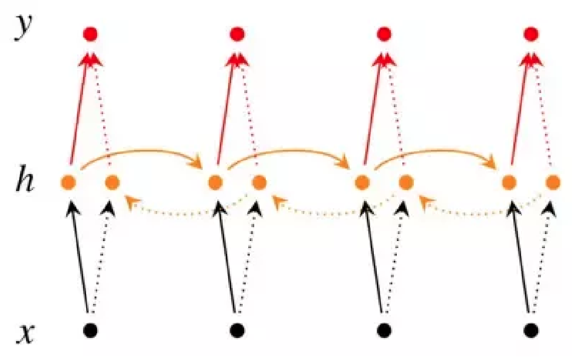

Figure 2: Bi-LSTM layer structure

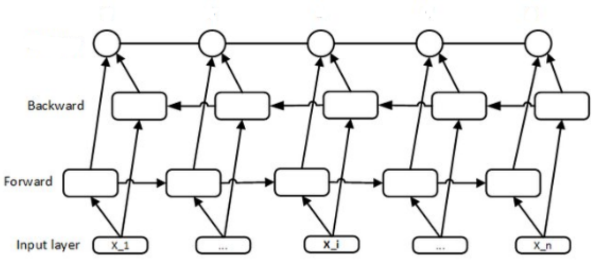

Figure 3: Bi-LSTM neural architecture

Parameters. We selected 20% of data as validation set to tune hyper-parameters. For example, the weight of the input layer in both model was initialized with 100 dimensional word vectors of the Glove word embedding. The training parameters setting as follows: using Skip-Gram model (cbow is 0), size is 100, window is 10, min count is 1. These neural network model we proposed were implemented in Keras. In the convolutional layer, for computational reasons, the length of every input for review was 100; the number of convolution filters was 64; the filter width was 4; and the activation function was ReLU. In the LSTM layer, the dimensions of the hidden states and cell states in the LSTM cells were both set to 100. In the classification output layer, we set 1 hidden unit with softmax activation function. The batch size of the neural network was 32, and the optimizer was Adam. In our experiment, all dropout values were set to 0.2. Shuffling was not performed after every epoch. The training procedure periodically evaluated the binary cross-entropy objective function on the training and validation sets. All setting is the same for the bidirectional LSTM model, and we set the merge mode for two layers as concatenate.

5 RESULTS & DISCUSSION

To compared with CNN-LSTM and the Bi-directional LSTM models, we used LSTM as the baseline method. In addition, during the experiments, we found that there are 45% tokens of our MOOC corpus missing in the GloVe model. So, in order to analyzed the impact of different word embedding weight pre-trained by GloVe and MOOC corpus, we kept the same deep learning model with Glove word embedding and MOOC corpus word embedding separately.

5.1 Research Question 1

Table 4 shows the accuracy of classifying whether a post is a question. We considered Support Vector Machine (SVM) as the baseline mode. From the results, we found that baseline SVM performed worst compared with all deep learning models. This indicated that deep learning has the potential of dealing with the complex text context for MOOC discussion forum.

5.2 Research Question 2

Table 5 - 6 shows the results of classifying question into three types. We observed that, in most cases, Bi-LSTM performed better than CNN-LSTM, and better than LSTM. One possible reason is that, Bi-LSTM capture both forward and backward text context information, which is very important in the MOOC specific discussion forum corpus. Also, though CNN with max-pool layer will lose dome information of sequence order, the adjacent features it obtained are more important than the words order in sentences.

5.3 Research Question 3

Table 4 - 6 shows the results of classifying question into three type. We found that with GloVe word embedding, the performance is

	2013				2015			
	SVM	LSTM	CNN-LSTM	Bi-LSTM	SVM	LSTM	CNN-LSTM	Bi-LSTM
Question-Tf-idf	0.16	NA	NA	NA	0.07	NA	NA	NA
Question-GloVe	NA	0.27	0.21	0.25	NA	0.25	0.2	0.2
Question-MOOC	NA	0.65	0.65	0.72	NA	0.64	0.58	0.75

Table 4: Question classification with MOOC corpus

	2013			2015		
	LSTM	CNN-LSTM	Bi-LSTM	LSTM	CNN-LSTM	Bi-LSTM
CQ	0.74	0.74	0.75	0.49	0.51	0.52
TQ	0.1	0.1	0.1	0.31	0.32	0.32
LQ	0.16	0.16	0.16	0.16	0.17	0.16

Table 5: Question type classification with GloVe

	2013			2015		
	LSTM	CNN-LSTM	Bi-LSTM	LSTM	CNN-LSTM	Bi-LSTM
CQ	0.6	0.63	0.74	0.73	0.75	0.85
TQ	0.72	0.75	0.9	0.5	0.72	0.84
LQ	0.84	0.89	0.9	0.83	0.88	0.9

Table 6: Question type classification with MOOC corpus

very pool compared with MOOC corpus embedding. One possible explanation is that, there are 45% of tokens missing in the GloVe model, so when deep learning model assign weights to words with is, many of the words weight information are missing. In addition, the GloVe trained based on Wikipedia 2014 corpus, the results also indicated that the text context is very different from MOOC discussion forum to Wikipedia.

6 CONCLUSION

In this paper, we investigated the framework to identify questions in the MOOC discussion forums and to classify questions into different type for building automatic question answering system in the future. We examined whether deep learning helps instructors to identify students posting questions in the MOOC discussion forum. Then, we proposed two deep learning structures which avoid heavy feature engineering compared to classical machine learning methods: a combination of LSTM and CNN, which take the advantage of LSTM that remember the past information for sequential prediction and of CNN that chose useful adjacent features to improve the performance and reduce the time cost; bi-directional LSTM which is able to capture both past and the future information. From the results, we concluded that among the 'BDE' MOOC dataset, the Bi-directional LSTM was the best model compared with CNN-LSTM and baseline LSTM, which indicated that when MOOC students seeking help, the long period of forward and backward information is more important than the adjacent surrounding information than only the forward information. In addition, among our dataset, word2vec pre-trained by the MOOC corpus performed much better than the Glove pre-trained for the MOOC discussion forum question classification task, might because the MOOC discussion forum context is very different from Wikipedia's.

In the future, we plan to use the classified question posts to developed a automatic question answering system, which can both suggest potential answers for teaching assistants and students who post the question. Also, there existed a cold start problem at the beginning of each semester, to address this, we will train cross semester classifiers in the future.

ACKNOWLEDGMENTS

This research was supported by NSF #1821475 "Concert: Coordinating Educational Interactions for Student Engagement" Collin F. Lynch, Tiffany Barnes, and Sarah Heckman (Co-PIs).

REFERENCES

[1] [n. d.]. When should one use bidirectional LSTM as opposed to normal LSTM? https://www.quora.com/When-should-one-use-bidirectional-LSTM-as-opposed-to-normal-LSTM. Accessed: 2010-09-30.

[2] Kenneth J Berry and Paul W Mielke Jr. 1988. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement* 48, 4 (1988), 921–933.

[3] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist* 49, 4 (2014), 219–243.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[5] Yi Cui, Wan Qi Jin, and Alyssa Friend Wise. 2017. Humans and machines together: Improving characterization of large scale online discussions through dynamic interrelated post and thread categorization (DIPTiC). In *4th Annual ACM Conference on Learning at Scale, L@ S 2017*. Association for Computing Machinery, Inc.

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*. 2121–2129.

[7] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. (1999).

[8] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 6645–6649.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[10] Dac-Viet Lai, Nguyen Truong Son, and Nguyen Le Minh. 2017. Deletion-based sentence compression using Bi-enc-dec LSTM. In *International Conference of the Pacific Association for Computational Linguistics*. Springer, 249–260.

[11] Fu-Ren Lin, Lu-Shih Hsieh, and Fu-Tai Chuang. 2009. Discovering genres of online discussion threads via text mining. *Computers & Education* 52, 2 (2009), 481–495.

[12] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742* (2017).

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[14] Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med* 37, 5 (2005), 360–363.

[15] Xu Wang, Miaomiao Wen, and Carolyn P Rosé. 2016. Towards triggering higher-order thinking behaviors in MOOCs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 398–407.

[16] Xu Wang, Diyi Yang, Miaomiao Wen, Kenneth Koedinger, and Carolyn P Rosé. 2015. Investigating How Student's Cognitive Behavior in MOOC Discussion Forums Affect Learning Gains. *International Educational Data Mining Society* (2015).

[17] Xiaocong Wei, Hongfei Lin, Liang Yang, and Yuhai Yu. 2017. A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification. *Information* 8, 3 (2017), 92.

[18] Miaomiao Wen, Diyi Yang, and Carolyn Rose. 2014. Sentiment Analysis in MOOC Discussion Forums: What does it tell us?. In *Educational data mining 2014*. Citeseer.

[19] Alyssa Friend Wise and Yi Cui. 2018. Unpacking the relationship between discussion forum participation and learning in MOOCs: content is key. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM, 330–339.

[20] Alyssa Friend Wise, Yi Cui, and Jovita Vytasek. 2016. Bringing order to chaos in MOOC discussion forums with content-related thread identification. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM, 188–197.

[21] Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *Icml*, Vol. 97. 412–420.

[22] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. 2015. A C-LSTM neural network for text classification. *arXiv preprint arXiv:1511.08630* (2015).

[23] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639* (2016).