A Multidimensional Deep Learner Model of Urgent Instructor Intervention Need in MOOC Forum Posts

Laila Alrajhi[1], Khulood Alharbi[1] and Alexandra I. Cristea[1]

[1] Computer Science, Durham University, Durham, UK

{laila.m.alrajhi, khulood.o.alharbi,
alexandra.i.cristea}@durham.ac.uk

Abstract. In recent years, massive open online courses (MOOCs) have become one of the most exciting innovations in e-learning environments. Thousands of learners around the world enroll on these online platforms to satisfy their learning needs (mostly) free of charge. However, despite the advantages MOOCs offer learners, dropout rates are high. Struggling learners often describe their feelings of confusion and need for help via forum posts. However, the often-huge numbers of posts on forums make it unlikely that instructors can respond to all learners and many of these urgent posts are overlooked or discarded. To overcome this, mining raw data for learners' posts may provide a helpful way of classifying posts where learners require urgent intervention from instructors, to help learners and reduce the current high dropout rates. In this paper we propose, a method based on correlations of different dimensions of learners' posts to determine the need for urgent intervention. Our initial statistical analysis found some interesting significant correlations between posts expressing sentiment, confusion, opinion, questions, and answers and the need for urgent intervention. Thus, we have developed a *multidimensional deep learner model* combining these features with natural language processing (NLP). To illustrate our method, we used a benchmark dataset of 29598 posts, from three different academic subject areas. The findings highlight that the combined, multidimensional features model is more effective than the text-only (NLP) analysis, showing that future models need to be optimised based on all these dimensions, when classifying urgent posts.

Keywords: MOOCs, Intelligent Tutoring System, Urgent Intervention, Deep Learning, Mixed Data.

1 Introduction

MOOCs are open distance-learning environments with large-scale enrolment [1]. Since their emergence as a popular mode of learning in 2012 [2], they have been delivering learning opportunities to a wide range of learners free or at low cost across different domains around the world [3], attracting thousands of learners to take advantage of the offered opportunities [4]. Amongst these, MOOC *online discussion forums* offer opportunities for learners to ask questions and express their feelings about course content and their learning progress, via *posts*. These can connect learners to learners, or learners to instructors. Instructor intervention is sought after, and could make the difference between a learners completing the course or not. However, due to

the large-scale participation in these platforms and extremely high ratios of learners to instructors, it is difficult for instructors to monitor all posts and determine when to intervene [5]. Therefore, researchers, MOOCs designers, and universities have begun to pay more attention to instructors' presence and their interventions in MOOC-based environments. As a result, many recent studies have focussed on detecting struggling learners' posts, to predict when they require intervention by instructors. Some of these approaches use features extracted from the properties of posts [6] and others are based on text-only features [7] [8] [9]. However, few studies have combined mixed data such as text data with metadata [10] [11], and they are limited, as they are all based on shallow machine learning (ML) only. Recently, deep learning models have been used for text-classification tasks [12].

Thus, we formulated the following two research questions:

RQ1: *Is there a relationship between the various dimensions of the learners' posts and their need for urgent instructor intervention?*

RQ2: *Does using several dimensions as features in addition to textual data increase the model's predictive power of the need for urgent instructor intervention, when using deep learning?*

In this paper, we contribute thus by answering the above questions via building a new classifier for this area, based on a *deep learning model that incorporates different dimensions of MOOC posts, i.e., numerical data in addition to textual data, to classify urgent posts*.

2 Related Work

2.1 Analysis in MOOCs

Recently, in the MOOC context, there have been significant efforts to study, analyse and evaluate different aspects of learners including sentiment [13], confusion [14] or need of urgent intervention [8], to improve the educational quality of MOOC environments and improve MOOCs' overall educational outcomes.

In terms of sentiment analysis, researchers have employed sentiment analysis for different purposes; for instance, they used it to predict attrition [15], performance and learning outcome [16], emotions [17] and dropout [13] by using different machine learning approaches. These methods include statistical analysis, shallow machine learning and deep neural networks. A growing number of researchers have studied confusion; [18] explored click patterns to identify the impact of confusion on learner dropout; [14] attempted to assist confused learners, by developing a tool that recommends relevant video clips to learners who had submitted posts that indicated learner confusion.

However, while all of these studies focus mainly on employing learner sentiment and confusion to achieve different goals, they do not exploit *sentiment and confusion indicators* to predict urgent instructor intervention. Therefore, our research seeks to use these aspects as a metadata to predict urgency posts.

2.2 Urgent Intervention in MOOCs

Detection of the need for urgent instructor intervention is arguably one of the most important issues in MOOC environments. The problem was first proposed and tackled [6] as a binary prediction task based on instructors' intervention histories. They [6] used traditional models (logistic regression [LR], the linear Markov chain model [LMCM], and the global chain model [GCM]). A follow-up study [10] proposed the use of L1-regularised logistic regression as a binary classifier. They [10] predicted when learners required intervention or not, by adding prior knowledge about the type of forum (thread) as a feature, in addition to linguistic features of posts. Another study [11], tried to build a generalised model, using different shallow ML models with linguistic features with metadata ('Up_count', 'Reads' and 'Post_type') - some extracted using NLP tools. In general, studies used as inputs for classification models either text-only data [5] [7] [8] [9] [19], or different *post*-specific features, such as linguistic features, other metadata [6], or a combination of textual data and *post* features [10] [11]. Moreover, they either used *traditional machine learning classifiers* [10] [11], or, more recently *transfer* [5] [7] and *deep learning* [8] [9] [19], as explored next. *Transfer learning*, as cross-domain classification was proposed [7] by training different traditional classifiers (support vector machine [SVM] and logistic regression) on three different dimensions (confusion, urgency, and sentiment), before validating them across different domains. The study [7] found low cross-domain classification accuracy, but mentioned that transfer learning should be given more attention. Moreover, this model is based on text-only data. A follow-up study [5] proposed a transfer learning framework based on *deep learning* (Convolution-LSTM [long short-term memory]) to predict different dimensions (confusion, urgency, and sentiment) in posts, using textual data only. This study is the first to apply deep learning in filtering posts, to predict which learners require urgent intervention. The following studies are all based on deep learning and used only textual data as an input to the model. [9] classified urgent posts with recurrent convolutional neural networks (RCNN), which use the *embedded information* of a current word, to capture contextual information. [8] proposed a hybrid character-word neural network based on *attention*, to identify posts that require urgent instructor intervention, also adding course information associated with a given *post* for contextualisation. [19] produced EduBERT as a pre-trained deep language model for learning analytics, trained on forum data from different online courses. They classified the urgency of instructor intervention as a text classification tasks, by fine-tuning EduBERT.

To the best of our knowledge, no studies have used *deep learning as an urgency-classifier model with mixed-input data*. In our study, we incorporated several different dimensions combining numerical data with textual data.

3 Methodology

We aim here to analyse combining several different dimensions with textual data, to predict posts where learners require urgent intervention in a MOOC environment.

3.1 Dataset

In this study, we used the Stanford MOOC benchmark posts dataset [14], which is available to academic researchers by request. It covers three different domain areas: education, humanities/sciences, and medicine, and contains 29,604 anonymised posts from 11 courses. Each post was manually labelled by three independent human coders to create a gold-standard dataset. Each post was evaluated against six categories/dimensions (*sentiment, confusion, urgency, opinion, question,* and *answer*). *Opinion, question* and *answer* were assigned binary values while *sentiment, confusion* and *urgency* were assigned values based on a scale of 1-7. To explain, for *sentiment,* 1 = extremely negative and 7 = extremely positive; for *confusion,* 1 = extremely knowledgeable and 7 = extremely confused; for *urgency,* 1 = no reason to read the post and 7 = extremely urgent: instructor definitely needs to reply. The final gold-standard dataset contains a column for each dimension, based on computing scores between coders. For more information about the coding process and the creation of the gold-standard dataset see their website [20]. Although the original dataset is multivalued, in order not to add additional complexity, we followed [8] and structured the problem of detecting urgent posts as a binary classification task by converting the (1-7) scale to binary values:

- Urgent intervention required > 4 $\Rightarrow$ Need for urgent intervention (1)
- Otherwise <= 4 $\Rightarrow$ No need for intervention (0)

We prepared the experimental data by excluding posts that contained, e.g. only numbers; this produced 29,598 'text' posts, where 23,992 were non-urgent posts (81%) and 5,606 urgent (≈19%). Next, we cleaned the noisy data, via removing automated anonymisation (e.g., <nameredac>, <phoneredaci>, <zipredaci>) and also, removing punctuation and hyperlinks, as in [5]. We also applied case-folding and lemmatisation [8]. However, we kept the stopwords, as recommended in [21] to improve accuracy.

3.2 Exploratory Statistical Analysis

To address the first research question, first, we calculated the relationship between the ratio number of non-urgent and urgent posts using the 5 dimensions (sentiment, confusion, opinion, question, and answer) for these posts. For the first two dimensions (sentiment and confusion), we rounded down the values to integers (e.g., 1 and 1.5 to 1; 2 and 2.5 to 2; etc.) merely for visualisations purposes. Then we calculated the mean value (μ) for the different aspects (the sentiment for non-urgent versus urgent posts; confusion with urgency and without; etc). To discover if the data were normally distributed, we applied the commonly used Kolmogorov-Smirnov (K-S) test. As the data were not normally distributed, we used a Mann-Whitney U test to check if the differences were significant. Then, we calculated the Bonferroni correction, as multiple comparisons were conducted. Finally, we measured (Pearson product-moment) correlations between non-urgent and urgent posts over the other dimensions. For correlation between non-urgent/urgent posts with sentiment and confusion values, we converted the scale to positive/negative: positive if the value was > 4 and negative otherwise.

3.3 Predictive Urgent Intervention Models

The first step towards answering the second research question was to develop a basic model based on text-only data and then incorporate other dimensions (sentiment scale, confusion scale, opinion value, question value and answer value) as numerical features. In general, we trained the text data (learners' posts) with a convolutional neural network (CNN) model and the numerical data (multiple dimensions) with a multi-layer perceptron (MLP) model (Fig. 1). We selected CNN to classify text by following [8], as they reported that TextCNN outperforms LSTM. Note though that our goal was to show the power of the multidimensional approach and not optimise the individual parts of our classifier.

Fig. 1. Different types of data with different networks.

We divided the data into two distinct sets: one for *training* and the other for *testing* (80% and 20%, respectively) using stratified sampling to ensure that the training and testing sets have approximately the same distribution of the different classes (non-urgent and urgent), although the dataset has a large number of non-urgent posts.

Text Model. As shown in Fig. 2, in the text model, the first layer is the input layer, with a maximum length = 200, as we padded out each post to a predetermined length (200 words) by following the current state of the art [8], to control the length of the input sequence to the model. Then, the embedding layer reused pre-trained word embeddings (Word2vec GoogleNews-vectors-negative300) and was fine-tuned during training. We selected (Word2vec) as the pre-trained model, as [8] showed that it outperformed Glove on classifying urgency tasks. Next, for the CNN layer, we applied 1D Convolution with (*128* filters, kernel size of *{3,4,5}* and Rectified Linear Unit 'ReLU' as activation function) as in [8], to derive interesting features, followed by 1D Global max pooling, to produce our features. Then, for the drop-out layer, we used a drop-out rate of 0.5 as in [8] to prevent overfitting. Then, the fully connected layer with the sigmoid as an activation function was used to classify the output I as: 1-needs urgent intervention or 0 – no intervention required:

$$I = \begin{cases} 1, & if > .5 \\ 0, & if \leq .5 \end{cases} \quad (1)$$

After constructing the model, we trained it using the Adam optimisation algorithm, as in [8]. We used binary cross-entropy as a loss function because our problems involve binary decisions, and we used the popular metrics of *accuracy* to measure performance. In addition, for a more comprehensive result and to deal with potential majority class bias, we calculated *precision*, *recall* and *F1-score* for each class.

Overall Model (Text Model + Other Dimensions Model). The overall model is a general model that contains mixed data to predict urgent posts. Here, we added numerical data as features in addition to text. As an initial study, we combined the text data with meta-data in one single model; however, the model's performance was unsatisfactory. As our model combines multiple inputs and mixed data, we therefore constructed two different sub-models (Fig. 2), with the first sub-model being the text-only model.

The second sub-model is a multi-layer perceptron (MLP) neural network, with 5 inputs that represent the 5 dimensions (*sentiment*, *confusion*, *opinion*, *question* and *answer*). Then we added these features one-by-one to the MLP model as single inputs (one dimension at a time) to check the individual effect of each particular dimension. The next layer is a hidden layer with 64 neurons. This is followed by a fully-connected layer with the sigmoid as an activation function to classify the posts as in the text model.

The outputs from these two sub-models were combined via *concatenation*, to construct the overall model. Finally, a fully connected layer with the sigmoid activation function was used at the end of the network to classify the output, as in the sub-models.

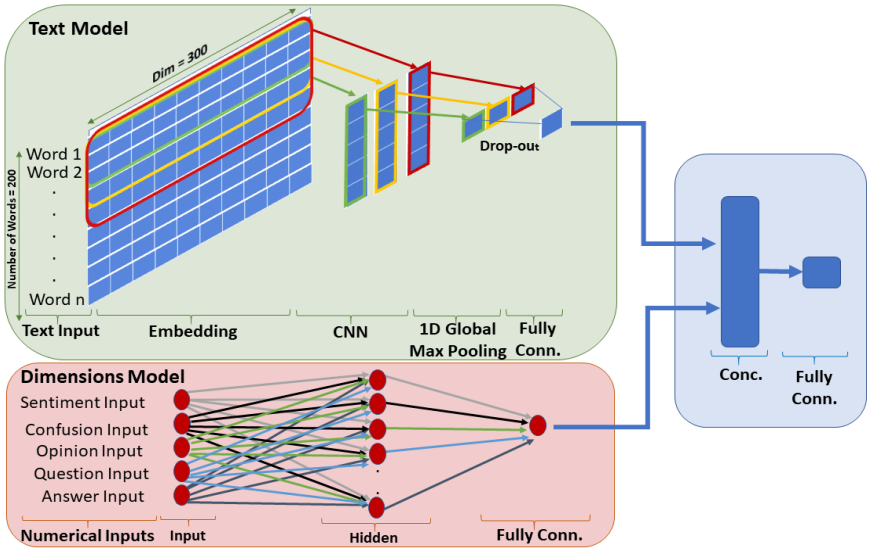

Fig. 2. Overall model.

After training, we applied McNemar's statistical hypothesis test to check if the observed differences between any two classifiers were statistically significant. We also applied the Bonferroni correction, to compensate for multiple comparisons.

4 Evaluation and Discussion

In this section, we present the charts and the results of the analysis of the relations between *non-urgent* and *urgent* posts with different dimensions, to address RQ1. Then, we review the results obtained after training each model to address RQ2.

4.1 Analysis

We analysed the relationship between the rates of non-urgent/urgent posts across the 5 different dimensions. As shown in Fig. 3 (left: Sentiment (1-7)), we observe that the number of urgent posts exceeds the number of non-urgent posts in the negative senti-

ment scale (1-3) and vice-versa: the number of urgent posts is less than that of non-urgent posts on the positive sentiment scale (5-7). We interpreted sentiment (4) as neutral. To reach this conclusion, we compared the values of (4) and (4.5) on the sentiment scale and found a higher proportion of non-urgent learners with a sentiment of (4.5). The figure also shows that for (right: Confusion (1-7)) the ratio of non-urgent posts is higher than that of urgent posts for non-confused posts, i.e. with confusion value between (1-3), in contrast to confused posts (5-7). We compared value (4) and (4.5) for confusion as well, and here, unlike for sentiment, results show a higher number of learners requiring urgent attention for the (4.5) value.

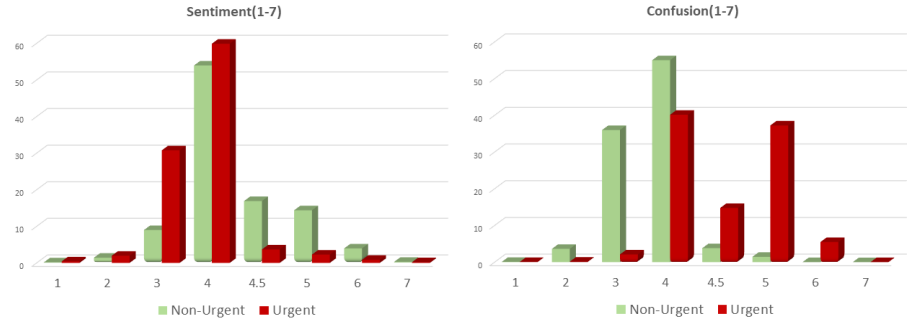

Fig. 3. The relationship between the ratio of the number of (non-urgent & urgent) posts and sentiment scale (1-7) (left), confusion scale (1-7) (right).

We performed a similar analysis for the remaining dimensions (*opinion*, *question* and *answer*), which are binary (Fig. 4). For *opinion*, most of the posts are non-urgent. For *question*, there are more urgent posts; this highlights that questions often represent posts where learners require urgent intervention. In *answer*, we found that, in general, most posts are not answered, indicating that most learners do not like to answer their peer's questions; this highlights the importance of instructor intervention. *Answer* posts, as expected normally represents non-urgent posts.

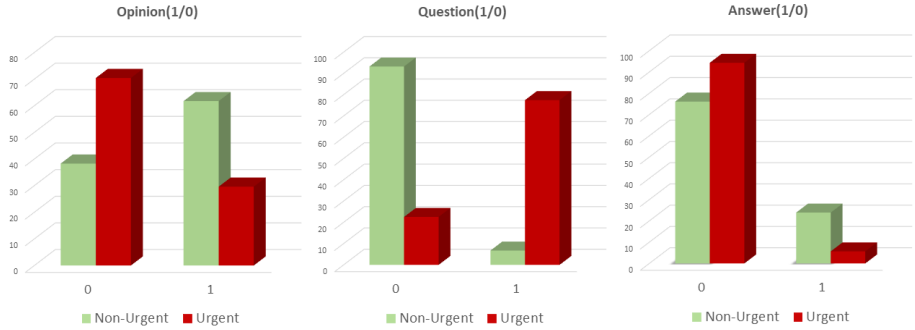

Fig. 4. The relationship between the ratio of the number of (non-urgent & urgent) posts and *opinion* (1/0) (left), *question* (1/0) (middle) and *answer* (1/0) right.

Next, we computed the averages on the sentiment dimension: the mean of the urgency sentiment was 3.83 and the mean of non-urgency sentiment was 4.25 (see Table 1). Importantly, this difference is statistically significant (Mann-Whitney U test: $p < 0.05$). Then, we repeated the same steps for all dimensions, as shown in Table 1. We

then applied a Bonferroni correction and found that $p < 0.01$, indicating that the set of all comparisons is significant.

Table 1. Average different dimensions with (non-urgent/urgent).

Dimension	Mean (non-urgent)	Mean (urgent)	P
Sentiment	4.25	3.83	$p < 0.01$
Confusion	3.75	4.59	$p < 0.01$
Opinion	0.61	0.29	$p < 0.01$
Question	0.06	0.77	$p < 0.01$
Answer	0.23	0.05	$p < 0.01$

Next, as explained in the methodology, we compared the dimensions. Correlation results are shown in Table 2, suggesting a strong correlation between *urgency* and *confusion* and also between *urgency* and *question*.

Table 2. Correlations between non-urgent/urgent posts reflected on different dimensions.

Dimension	Non-urgent/urgent
Sentiment	-0.244
Confusion	0.571
Opinion	-0.253
Question	0.691
Answer	-0.177

4.2 Predictive Intervention Models

Table 3 reports the performance of every trained model, as a comparison between different inputs. We calculated the average accuracy (*Acc*) and Precision (*P*), Recall (*R*) and F1-score (*F1*) per every class (0 as non-urgent) and (1 as urgent). The results revealed that adding all features as other dimensions (sentiment scale, confusion scale, opinion value, question value and answer value) in addition to texts increases classifier performance for classifying urgent posts.

Table 3. The performance results for different inputs (Acc,P,R,F1 %).

Inputs	Acc	Non-urgent (0)			Urgent (1)		
		P	R	F1	P	R	F1
Text	.878	.90	.95	.93	.73	.56	.64
Text + all features	**.912**	**.93**	.97	**.95**	.84	**.67**	**.74**
Text + sentiment	.879	.91	.95	.93	.73	.57	.64
Text + confusion	.872	.90	.95	.92	.73	.52	.61
Text + opinion	.874	.90	.95	.92	.71	.57	.63
Text + question	.903	.91	**.98**	.94	**.86**	.59	.70
Text + answer	.888	.92	.95	.93	.73	.64	.69

Next, we checked if these differences were statically significant (McNemar's test: $p < 0.05$) as shown in Table 4 ($\surd$ indicates a statistically significant difference in the disagreements between the two models while $\times$ signifies a statistically non-significant difference in the disagreements between the two models). The results confirm that there are differences between the (text + all features) model and the other models as they have different proportions of errors. Then we used a Bonferroni correction be-

tween (text + all features) and different models; we found that $p < 0.008$, meaning the set of all comparisons is significant.

Table 4. McNemar's test results between models.

	Text	Text+all features	Text+ sentiment	Text+ confusion	Text+ opinion	Text+ question	Text+ answer
Text							
Text+all features	√						
Text+ sentiment	×	√					
Text+ confusion	×	√	√				
Text+ opinion	×	√	×	×			
Text+ question	√	√	√	√	√		
Text+ answer	√	√	√	√	√	√	

5 Conclusion

Identifying when instructors should offer learner intervention is an extremely important issue in MOOC environments. In this paper, we have tackled this problem for the first time, as a multidimensional *post*-based learner model, exploring deep learning. Specifically, we compare text-based models with enriched models with the dimensions of (*sentiment, confusion, opinion, question* and *answer*). We also observed the relationship between urgent post rates and these dimensions. We showed that learners' negative feelings, misunderstandings, lack of desire to express an opinion, number of questions, and decreasing number of answers increase for learners in need of urgent intervention, possibly due to the nature of people. Our contributions include showing that adding these dimensions as features, in addition to text, leads to better predictive performance in deep learning models. Moreover, we constructed a new architecture based on sub-models to train this multidimensional, mixed data.

References

1. Arguello, J. and K. Shaffer. *Predicting speech acts in MOOC forum posts*. in *Ninth International AAAI Conference on Web and Social Media*. 2015.
2. Yan, W., et al. *Exploring Learner Engagement Patterns in Teach-Outs Using Topic, Sentiment and On-topicness to Reflect on Pedagogy*. in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. 2019. ACM.
3. Gupta, R. and N. Sambyal, *An understanding approach towards MOOCs.* International Journal of Emerging Technology and Advanced Engineering, 2013. **3**(6): p. 312-315.
4. Drake, J.R., M. O'Hara, and E. Seeman, *Five principles for MOOC design: With a case study.* Journal of Information Technology Education: Innovations in Practice, 2015. **14**(14): p. 125-143.
5. Wei, X., et al., *A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification.* Information, 2017. **8**(3): p. 92.

6. Chaturvedi, S., D. Goldwasser, and H. Daumé III. *Predicting instructor's intervention in MOOC forums*. in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014.
7. Bakharia, A. *Towards cross-domain mooc forum post classification*. in *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. 2016. ACM.
8. Guo, S.X., et al., *Attention-Based Character-Word Hybrid Neural Networks with semantic and structural information for identifying of urgent posts in MOOC discussion forums.* IEEE Access, 2019. **7**: p. 120522-120532.
9. Sun, X., et al. *Identification of urgent posts in MOOC discussion forums using an improved RCNN*. in *2019 IEEE World Conference on Engineering Education (EDUNINE)*. 2019. IEEE.
10. Chandrasekaran, M.K., et al., *Learning instructor intervention from mooc forums: Early results and issues.* arXiv preprint arXiv:1504.07206, 2015.
11. Almatrafi, O., A. Johri, and H. Rangwala, *Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums.* Computers & Education, 2018. **118**: p. 1-9.
12. Zhou, C., et al., *A C-LSTM neural network for text classification.* arXiv preprint arXiv:1511.08630, 2015.
13. Wen, M., D. Yang, and C. Rose. *Sentiment Analysis in MOOC Discussion Forums: What does it tell us?* in *Educational data mining 2014*. 2014. Citeseer.
14. Agrawal, A., et al. *YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips*. in *the 8th Intl. Conference on Educational Data Mining*. 2015.
15. Chaplot, D.S., E. Rhim, and J. Kim. *Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks*. in *AIED Workshops*. 2015.
16. Tucker, C., B.K. Pursel, and A. Divinsky, *Mining student-generated textual data in MOOCs and quantifying their effects on student performance and learning outcomes.* The ASEE Computers in Education (CoED) Journal, 2014. **5**(4): p. 84.
17. Moreno-Marcos, P.M., et al. *Sentiment Analysis in MOOCs: A case study*. in *2018 IEEE Global Engineering Education Conference (EDUCON)*. 2018. IEEE.
18. Yang, D., et al. *Exploring the effect of confusion in discussion forums of massive open online courses*. in *Proceedings of the second (2015) ACM conference on learning@ scale*. 2015. ACM.
19. Clavié, B. and K. Gal, *EduBERT: Pretrained Deep Language Models for Learning Analytics.* arXiv preprint arXiv:1912.00690, 2019.
20. Agrawal, A. and A. Paepcke. *The Stanford MOOCPosts Data Set*. Available from: https://datastage.stanford.edu/StanfordMoocPosts/.
21. Wise, A.F., et al., *Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling.* The Internet and Higher Education, 2017. **32**: p. 11-28.