A Good Classifier is Not Enough: a XAI Approach for Urgent Instructor-Intervention Models in MOOCs

Laila Alrajhi[1,4], Filipe Dwan Pereira[2], Alexandra I. Cristea[1] and Tahani Aljohani[3]

[1] Computer Science, Durham University, Durham, UK
{laila.m.alrajhi, alexandra.i.cristea}@durham.ac.uk
[2] Computer Science, Federal University of Roraima, Boa Vista, Brazil
filipe.dwan@ufrr.br
[3] taljohani7@gmail.com
[4] Educational Technology, King Abdulaziz University, Jeddah, SA

Abstract. Deciding upon instructor intervention based on learners' comments that need an urgent response in MOOC environments is a known challenge. The best solutions proposed used automatic machine learning (ML) models to predict the urgency. These are 'black-box'-es, with results opaque to humans. EXplainable artificial intelligence (XAI) is aiming to understand these, to enhance trust in artificial intelligence (AI)-based decision-making. We propose to apply XAI techniques to interpret a MOOC intervention model, by analysing learner comments. We show how pairing a good predictor with XAI results and especially colour-coded visualisation could be used to support instructors making decisions on urgent intervention.

Keywords: MOOCs, comments, urgent intervention, NLP, XAI.

1 Introduction

Instructor intervention in MOOCs may reduce the problem of learner dropout, as it has recently been proven that learners who need intervention are less likely to complete the course (only 13%) [1]. Recently, intervention in MOOC attracted growing interest from researchers, to help instructors in interventions based on learners' comments [2] [3] [4] [5]. Intervention systems classified the learner comments into two categories: urgent and non-urgent [6]. Although these systems need to be accurate in their decisions, it is difficult to achieve this, as urgency decisions are hard to make, even for a human [7].

This work deals with the intervention problem. Our initial goal is the proof of concept of using explainable AI for this task of urgent intervention, as this had not been done before. For understanding 'How' and 'Why' the model decisions are made, we explained thus not only the intervention model prediction, but also compared it with human decision making. We formalise our research question as:

RQ: How to construct a transparent XAI model to detect urgent intervention towards supporting instructors' decisions?

In terms of the contribution, to the best of our knowledge *this is the first time that text classification explainability has been applied to an instructor intervention model*.

2 Methods

Our research consists of three basic stages as follows: first, construct an 'urgent' gold-standard dataset, via human experts annotating comments (Section 2.1). Next, build an automatic urgent intervention model via BERT (Section 2.2). Then, explain the model and visualise words importance, to understand the decision (Section 2.3).

2.1 Constructing the Gold-standard Dataset

We collect and prepare our benchmark corpus, as a case study, based on real-world data from the FutureLearn MOOC environment platform, here, the 'Big data' course, conducted during 2016 – selected as being a topic of current interest for learning; additionally, we expected it to contain many urgent cases, as being (arguably) a more challenging topic. The dataset consists of learner comment texts ('posts') and other features collected from the first 5 weeks (≈ 50%) of the 9-weeks-long course, to capture the comments that need intervention before dropout. We obtain thus 5786 comments, which, taking into account the hardship of the following manual annotation, were considered sufficient for the current task.

We thus manually annotate these comments, using three human domain experts and one author of this paper, following Agrawal et al.'s instructions [8]. We labelled urgency for every learner comment, mapped onto a scale (1-7) representing the range of urgency level (not urgent – extremely urgent). For validation, we calculated Krippendorff's α agreement value between all annotators, and we found the results very low between any subgroups (confirming prior research [7]). To address this problem, we decide to further convert the scale into a simpler, binary one (mapping 1:3 → 0, and 4:7 → 1). To be able to increase the reliability, we additionally dropped the annotator who disagreed strongly with other annotators. From the remaining three annotators, we calculate the label value, via the voting technique, since voting is the most common way to gather different opinions for the same task [9]. This result in a class size of ('0' non-urgent → 4903, '1' urgent → 883).

2.2 Fine-tuning the BERT model

As the preprocessing step, we split the data into training and testing sets, using the stratify method [10], to preserve the percentage of samples for each class; with the proportion of 80% training and 20% testing. Thus, the distribution of the training set is (0: 3922, 1: 706) and testing set is (0: 981, 1: 177). Then, for the training set, we split again, as 90% will be used for training, and 10% will be used for validation.

We fine-tune BERT, without any engineering features. We use the 'bert-base-uncased' version. Next, we prepare the text input, with the fixed maximum length 365, which is the maximum number of words on all comments; this will pad all comments to the maximum length. Then we train the model, by defining batch size = 8, number of training epochs = 4 and AdamW, as optimiser, with learning rate = 2e-5. Finally, we evaluate the prediction model performance on the test set, and save the pre-trained model, to use it later for the interpreting.

2.3 Interpreting the BERT Model

After training the model, we interpret our BERT model, by using the Captum package, which supports classification models. We interpret it via the BertForSequenceClassification in Captum from Captum_BERT colab [11], by creating the Layer Integrated Gradients explainer, to identify which words have the highest attribution to the model's output. To illustrate how to use our method and reply to RQ, we randomly choose a single comment and visualise the explainability results with the attribution score and highlight the word importance.

3 Results

The results obtained from BERT to predict the urgent comments show that the accuracy score is high (0.92). However, as the data is extremely unbalanced, we use additional metrics to evaluate the classifier (precision, recall and F1-score) for every class, see Table 1. Please note that here, whilst working with a decent classifier, our focus is not on the optimisation of the classifier, but on the explanation of the obtained results.

Table 1. The results of the BERT classifier.

	Precision	Recall	F1-score
0	.95	.95	.95
1	.73	.71	.72

As previously mentioned, our goal is to analyse the learner comments and explain the text classification decision using Captum, to understand the reasons behind the predictions. Here we chose a random comment prediction from the test set, then show the explainability results, with highlighted text, as shown in Fig. 1. The attribution score = 1.45 and the different colours reflect the effect of word attribution towards the prediction; and the level of highlighting depicts the importance of the feature, for the classification. Specifically, the green colour means a positive contribution (got, looking, understanding, be, …), whilst red contributes by decreasing the prediction score (forward, useful, …). In the case of the example below, we found that the predicted label is non-urgent (0) and the true label is also non-urgent (0). Such visualisation can further be used by an instructor to understand the decisions and recommendations of a classifier for urgency detection in learners' chats on MOOCs.

Fig. 1. Screenshots of Captum explanations.

4 Conclusion

The objective of this paper was to provide an explanation of the machine learning decision, for a specific text classification problem, that of explaining individual predictions in the urgent intervention task in a MOOC environment.

Here, this work also represents a proof-of-concept of using explainable AI on imbalanced data. Moreover, we advance the field of urgency prediction, proposing a method for potentially supporting instructor intervention.

References

1. Alrajhi, L., et al. *Urgency Analysis of Learners' Comments: An Automated Intervention Priority Model for MOOC*. in *International Conference on Intelligent Tutoring Systems*. 2021. Springer.
2. 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.
3. 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.
4. Alrajhi, L., K. Alharbi, and A.I. Cristea. *A Multidimensional Deep Learner Model of Urgent Instructor Intervention Need in MOOC Forum Posts*. in *International Conference on Intelligent Tutoring Systems*. 2020. Springer.
5. Khodeir, N.A., *Bi-GRU Urgent Classification for MOOC Discussion Forums Based on BERT.* IEEE Access, 2021. **9**: p. 58243-58255.
6. 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.
7. Chandrasekaran, M.K., et al., *Learning instructor intervention from mooc forums: Early results and issues.* arXiv preprint arXiv:1504.07206, 2015.
8. Agrawal, A. and A. Paepcke. *The Stanford MOOCPosts Data Set*. Available from: https://datastage.stanford.edu/StanfordMoocPosts/.
9. Troyano, J.A., et al. *Named entity recognition through corpus transformation and system combination*. in *International Conference on Natural Language Processing (in Spain)*. 2004. Springer.
10. Farias, F., T. Ludermir, and C. Bastos-Filho, *Similarity Based Stratified Splitting: an approach to train better classifiers.* arXiv preprint arXiv:2010.06099, 2020.
11. Captum. *Captum_BERT*. 2022; Available from: https://colab.research.google.com/drive/1pgAbzUF2SzF0BdFtGpJbZPWUOhFxT2NZ.