Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Automatic Classification of MOOC Forum Messages to Measure the Quality of Peer Interaction

Urvi SHAH[a*], Richa RAMBHIA[a], Prakruti KOTHARI[a], Rekha RAMESH[a] & GargiBANERJEE[b]
[a]*Department of Computer Engineering, Shah & Anchor Kutchhi Engineering College, University of Mumbai, India*
[b]*Educational Technology Programme, IIT Bombay, India*
*umshah99@gmail.com

Abstract: Discussion forum is an integral part of many MOOCs as it provides a platform for peer interaction among learners. The quality of peer interaction is an indicator of the potential for peer learning. Thus, quality of peer interaction provides instructors with an actionable insight into the extent of critical or higher level thinking that learners are engaged in and is a measure of the learning effectiveness of the course. It is daunting for instructors to manually analyze the forum messages to gain this insight. To address this issue, we attempted to develop a system for automatic classification of forum messages that will inform instructors on the quality of peer interaction happening in the forum. Our system classifies messages into predefined classes based on the Interaction Analysis Model phases. We explored and implemented multiple machine learning models. A general accuracy of 95%-97% was observed among the models and no model outperformed the other by a great margin. The needfor such a system has become all the more relevant in the current Covid-19 pandemic situation, where all physical classrooms have had to migrate to an online setting.

Keywords: Massive open online courses; discussion forum; peer interaction; automatic classification; machine learning; neural network

1. Introduction

Discussion forums (DF) offer a platform for asynchronous communications that facilitates interactions and communications among learners and instructors, and it also helps learners build a community within the MOOC (Wong et.al, 2015). Analysis of forum messages is critical to instructors to obtain information about the quality of peer interaction taking place in the MOOC like a measure of learner engagement in higher-order thinking since the instructor has no face-to-face interaction with learners. Manual analysis of learners' posts is time consuming. Automatic classification can provide the requisite information to instructors enabling them to adopt measures to increase such type of interactions among learners in the remaining part or the next run of the course.

The existing work on message classifiers for the forum focuses on identifying messages for instructors that need their intervention. While they are useful, they do not provide a holistic view of the peer interaction happening in the forum. Hence, we developed a system that automatically classifies messages, helping the instructor capture and analyze the quality of peer interaction in DFs and the potential learning thereof. The forum data from two separate runs of an xMOOC, offered by the Indian Institute of Technology Bombay via IITBombayX platform, was used. The system automates the process of classification of forum messages into predefined classes that indicate the quality of peer interaction, based on the Interaction Analysis Model (IAM) (Gunawardena et.al, 1997). We explored multiple machine learning models such as decision tree- based models, rule-based models, statistical models, SVM, distance-based classifiers and deep learning models with the application of various pre-processing variants and analyzed their performances. The proposed system will be useful to instructors who have had to migrate to online teaching in the current Covid- 19

pandemic situation.

2. Related Work

We surveyed the existing literature on the importance of DF analysis and the machine learning based classifiers. Peer interaction fosters the 'highest level of collaboration' and critical reflection among learners (Toven- Lindsey et.al, 2015). MOOC forums often witness an overwhelming volume of orphaned posts that causes scatter (Manathunga et.al, 2017; McGuire, 2013), leading to cognitive overload for learners and limits peer interaction (Tawfik et.al, 2017).

Ntourmas et.al (2021) and Ntourmas et.al (2019) address the issue of information overload by developing supervised classification models to assist instructors in detecting forum discussions that need their intervention. Wise et.al (2016) addresses the issue of scatter and information overload in MOOC DFs by developing a model that categorizes and identifies threads based on their relation to the course content. In a similar work, Yi Cui and Alyssa Friend Wise (2015) investigated the extent to which the learner asked and the instructor answered content based questions in MOOC forums by building a classification model using linguistic features extracted by Lightside Researcher's Workbench. Yiqiao Xu and Collin F. Lynch (2018) proposed an identification framework to identify and analyse help-seeking post.

The existing work on analysis of MOOC forum messages address the challenges instructors and learners face in locating the relevant set of messages from the overwhelming volume of forum messages and the resultant scatter. We found that the classification models that are built focuses on identifying messages posted by learners that instructors need to address. None of them aim to capture the quality of peer interaction that would aid instructors to estimate the amount of higher order thinking that is happening through the forum. This quality analysis is important feedback for instructors on the learning effectiveness of the course but it is extremely difficult to gauge manually. Hence, there is a need for automation of analysis of forum messages to measure the quality of peer interaction. We built a system that will automatically classify the messages in terms of the thinking level expressed in the messages. The system output will provide instructors with actionable insight into the extent of critical thinking occurring among the learners.

3. Methodology

In this section, we mention details about the dataset used and proposed system to be built in order to bridge the gap present in current work.

3.1 Context

We took the forum data from two separate runs of the same teacher professional development MOOC. We refer to the two runs as T1 (first run) and T2 (second run) in this paper. The teacher-learners were from multiple domains like Mathematics, English, Computer Science, Science and Social Science. The DF consists of Comment-thread (initialization of the post/discussion), Replies (responses to the comment thread) and Comments (responses to the replies) (Wong et.al, 2015). In this study, we consider each message as the unit of analysis, irrespective of their initiation source. Permission to use forum data for research purposes was obtained from the learners through IITBombayX.

3.2 Data Description

In this section we discuss data extraction, the thematic analysis and the nature of data observed.

3.2.1 Data Extraction

The prerequisite for the labelling phase was to extract out the relevant attributes from the data in the

T1 and T2 dataset based on a set of rules. The attributes are Comment_count, Comment thread_id, Parent_id and Sk_id. Comment_count is the number of comment replies in this thread. This includes all responses and replies, but does not include the original post that started the thread. Comment thread_id specifies the id the Comment Thread to which a specific Comment belongs. The Parent_id is the id of the response-level Comment that this Comment is a reply to. Sk_id if null, the type is Comment Thread, else type is Comment.

3.2.2 Thematic Analysis

The classification of our messages is based on the IAM (Gunawardena et.al, 1997). In our system, we followed the below classification:

1. Superficial: Shallow messages with replies of greeting type from which learning is difficult to infer.
2. IAM phases: Messages that are elaborated in five phases (Gunawardena et.al, 1997).
    a. Sharing or Comparing of Information: Statement of observation or opinion, agreement or asking and answering questions to clarify details of statement, definition, description.
    b. Discovery of dissonance and inconsistency: Statements of disagreement, asking and answering questions to clarify the source and extent of disagreement.
    c. Negotiation of Meaning or Co-construction of knowledge: Clarification of the meaning of terms. Negotiation of the relative weight to be assigned to arguments, identification of areas of agreement or overlap among conflicting concepts.
    d. Testing and modification of proposed synthesis: Testing the proposed synthesis against 'received fact' as shared by the participants and/or their culture, testing against existing cognitive schema, testing against personal experience or formal data collected.
    e. Agreement/application of newly constructed meaning: Summarization of agreement(s), applications of new knowledge and metacognitive statements by the participants illustrating their (cognitive schema) has changed as a result of the interaction.
3. Off-topic: Content does not address any aspect of the current topic of discussion.

In this paper we referred to the above-mentioned a, b, c, as Sharing and Comparing, Dissonance, Negotiation and Co-construction. One of the authors did deductive coding of the T1 and T2 message rows using the mentioned categories as the codebook and following a set of general guidelines. For the Guided messages, labelling was done with respect to Focus Question. There were 6 Focus Questions in total, one for each week which were created by instructors to anchor discussions. The relevance of the Comment Threads is to be checked with respect to its Focus Question. For Non-Guided messages, since there are no focus questions, Comment Threads are to be judged considering the general nature of the sentence. Another author coded the same set of messages independently. Inter-rater reliability of the coding was established through discussion between the two coders till a complete consensus was achieved. There were 29355 data rows of T1 and 4707 data rows of T2. All 4707 of T2 data rowswere tagged. Since manual coding is labor intensive and time consuming, only 4039 data rows from T1 were arbitrarily picked and labelled.

3.2.3 Nature of Data

All messages including the comment threads, comments and replies were given equal importance and considered independently. The dataset contained two fields, namely message and category. Message refers to thread, comment, replies and the Category is obtained from coding the messages as mentioned in the above section. We proposed to build a model that aims to successfully classify the messages from the DF. We experimented with various classification approaches and compared them to check which would suit the best. We implemented machine learning algorithms and Keras Neural networks for categorizing the messages.

Superficial and Off-topic categories were not considered while building the classifier model since they are unlikely to contribute to learning. Only three phases of IAM were considered i.e.,

Sharing and Comparing, Dissonance, Negotiation and Co-construction because the remaining phases of the IAM were not found in the data. We found the data to be highly biased towards Sharing and Comparing with the majority of messages belonging to this phase. This is a common finding for MOOC forums where quality of learning has been found to be confined to the lowest level of critical thinking i.e., Sharing and Comparing information (Tawfik et.al, 2017). Consequently, we divided the IAM phases into two classes i.e., 'Sharing and Comparing' and 'Higher order thinking (HOT)' where HOT covers the IAM phases beyond 'Sharing and Comparing'. For the forum being analyzed, HOT consisted of messages belonging to two such IAM phases – Dissonance, Negotiation and Co-construction. The resultant distribution among the categories is highlighted in the figure below.

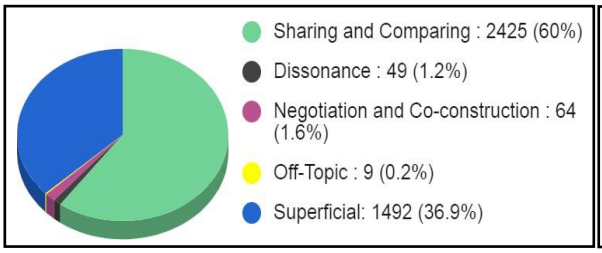

Figure 1. Distribution of Message Classes in T1.

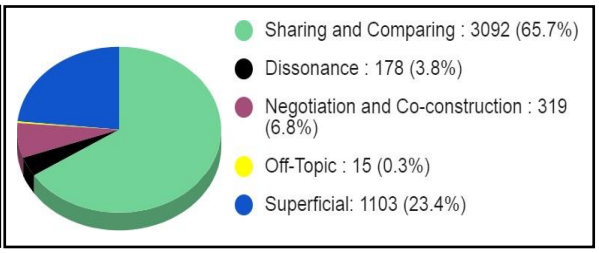

Figure 2. Distribution of Message Classes in T2.

3.3 Data Pre-processing

Our preprocessing consists of - tokenization, lowercasing, expansion of contractions, spell check, punctuation removal and custom stopword removal. In our work, we expanded contractions such as don't and can't to "do not" and "can not" in order to standardize the text. We created a list of custom stop words and removed them from the data. This is because, the predefined stop words in the English language nltk corpus contains words that are useful for our classification. Various combinations of the aforementioned preprocessing steps were carried out. The resultant data obtained in each case was fed into the models that were built. The three variants of pre-processing applied were - Variant 1: lowercasing, punctuation removal, Variant 2: lowercasing, punctuation removal, custom stopword removal and Variant 3: lowercasing, contraction removal, spell check, punctuation removal, custom stopword removal.

3.4 Proposed System

The below diagram illustrates the workflow that was followed for making the system.

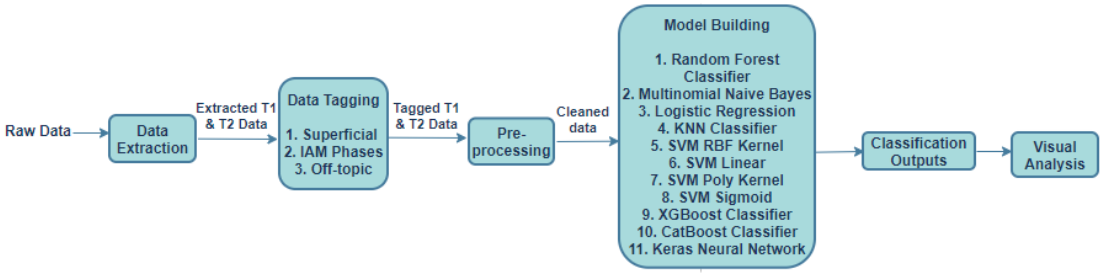

Figure 3. System Workflow.

4. Building and Training Models

The labeled messages of T1 and T2 were combined and 80% of these messages were arbitrarily picked and used for training the models. The remaining messages were used for testing the models. The three pre-processing combinations i.e., Variant 1, Variant 2 and Variant 3 were applied on these messages which were then used to build models for automated classification of the messages.

The built models were Random Forest Classifier, Multinomial Naive Bayes, Logistic

Regression, KNN Classifier, SVM RBF Kernel, SVM Linear, SVM Poly Kernel, SVM Sigmoid, XGBoost Classifier, CatBoost Classifier, and Keras Neural Network. A pipeline approach was used for fitting the machine learning models on the training data. In the case of all these models, the variants of unigrams, bigrams and unigrams & bigrams both were implemented. Tf-idf was used as the feature to transform the input data where these variants were passed as parameters. In the case of Keras Neural network, the same two class classification approach was used i.e., labelling of messages as either 'Sharing & Comparing' or 'HOT'. The pre-processing steps mentioned in all the aforementioned variants were carried out on the data. Four modes were used to transform the words into features in the case of Keras Neural Network. They are binary, count, tf-idf and others.

5. Results and Discussion

The below shown figure displays the performance measure of the implemented classification models. All the types and variants for each model were implemented but a few were picked for illustration as shown below. For example, let's consider the output of the XGBoost Classifier using unigram and variant 1 of preprocessing. HOT refers to the Higher Order Thinking classes and SC refers to the Sharing and Comparing class. 15 HOT messages and 784 SC were correctly classified into their respective classes whereas 21 HOT messages were classified as SC and 2 SC messages were classified as HOT.

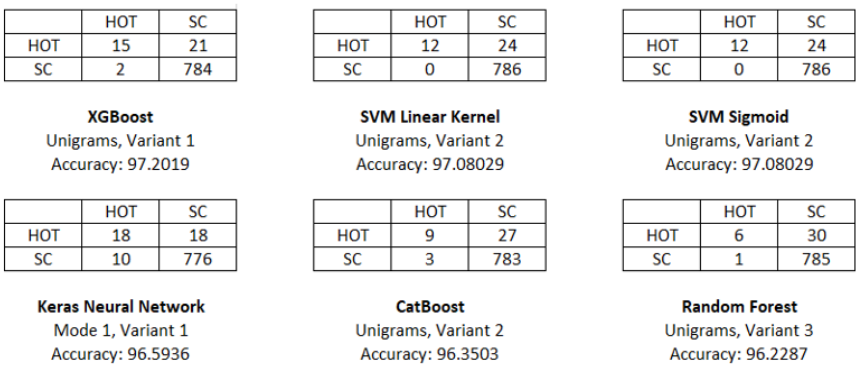

Figure 4. Performance Measure of Classification Models.

The general accuracy observed among the models was in the range of 95%-97%. Similar results were observed and no model outperformed the other by a great margin. As per our observations on the data in context, we found that the SVM Linear Kernel and XGBoost, both using unigrams, gave a relatively high accuracy of about 97% when it came to the overall classification of messages. Apart from SVM Linear Kernel and XGBoost, SVM Sigmoid Kernel model using unigrams, CatBoost using unigrams and Keras Neural Network gave satisfactory good results of around 96%. There were certain cases where the models were unable to classify any message belonging to the HOT category. This may be due to the disproportionate amount of data of the Sharing and Comparing class as compared to the HOT class. Models like KNN and Multinomial Naive Bayes were unable to classify 'Higher order thinking' type of messages correctly in all the considered combinations. We noticed a substantial increase in the accuracy on incrementing the training data that was fed to the models with the exception of some. SVM Linear Kernel and XGBoost gave the highest accuracy overall.

6. Conclusion

The objective of this study was to build an automatic classification system of MOOC forum messages to analyze the quality of peer interaction. We worked on the forum data from two separate runs of an xMOOC that ran on the IITBombayX platform for training and testing our system. We created a list

of custom stop words and removed them from the data. Multiple machine learning models were explored and implemented in order to classify forum messages into predefined classes based on the IAM coding scheme. A general accuracy of 95%-97% was observed among the models.

Our system will enable instructors to gauge the learning effectiveness of their course by automatically analyzing and classifying the messages in the course forum into 'Sharing and comparing' and 'Higher order thinking'. The distribution of the messages into these two classes will provide instructors with an actionable insight into the extent of peer learning occurring in the forum. This insight is crucial as it gives an estimate of the higher order thinking ensuing in the form that is likely to lead to effective learning. In absence of face-to-face interaction in a MOOC course, such insight provides the instructors an estimate of the learning happening without having to wait for the assessment grades of the learners. It enables instructors to decide when to intervene to increase productive interactions among learners in the DF during the progress of the course.

Our classification model can be effectively generalized to smaller online courses. This is significant given the current covid-19 pandemic scenario where instructors have had to migrate to online teaching. As part of future work, we plan to build a rule-based architecture on top of the existing systemor apply deeper NLP techniques to process the messages. The system's efficiency could be increasedif the data was non biased or with more messages belonging to higher phases of IAM beyond 'Sharingand Comparing'.

Acknowledgements

Authors would like to acknowledge IITBombayX and Education Technology program at IIT Bombay for their support and permission to use their course data for this study and Nexteducation Research Lab (IITB) who were co-sponsor of these courses.

References

Cui, Y., & Wise, A. F. (2015, March). Identifying content-related threads in MOOC discussion forums. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 299-303).

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). 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.

Manathunga, K., Hernández-Leo, D., & Sharples, M. (2017, May). A Social learning space grid for MOOCs: exploring a FutureLearn case. In *European Conference on Massive Open Online Courses* (pp. 243-253). Springer, Cham.

McGuire, R. (2013). Building a sense of community in MOOCs. Campus Technology, August, 31-33. Retrieved April 2, 2014, from the Campus Technology

Ntourmas, A., Avouris, N., Daskalaki, S., & Dimitriadis, Y. (2019, July). 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)* (pp. 1-8). IEEE.

Ntourmas, A., Daskalaki, S., Dimitriadis, Y., & Avouris, N. (2021). Classifying MOOC forum posts using corpora semantic similarities: a study on transferability across different courses. *Neural Computing and Applications*, 1-15.

Tawfik, A. A., Reeves, T. D., Stich, A. E., Gill, A., Hong, C., McDade, J., ... & Giabbanelli, P. J. (2017). The nature and level of learner–learner interaction in a chemistry massive open online course (MOOC). *Journal of Computing in Higher Education*, *29*(3), 411-431.

Toven-Lindsey, B., Rhoads, R. A., & Lozano, J. B. (2015). Virtually unlimited classrooms: Pedagogical practices in massive open online courses. *The internet and higher education*, *24*, 1-12.

Wise, A. F., Cui, Y., & Vytasek, J. (2016, April). Bringing order to chaos in MOOC discussion forums with content-related thread identification. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 188-197).

Wong, J. S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015, March). An analysis of MOOC discussion forum interactions from the most active users. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 452-457). Springer, Cham.

Xu, Y., & Lynch, C. F. (2018). 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* (pp. 1-6).