Computers & Education 118 (2018) 1–9

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums

Omaima Almatrafi*, Aditya Johri, Huzefa Rangwala

Department of Information Sciences and Technology, George Mason University, Fairfax, USA

ARTICLE INFO

Keywords:
Computer-mediated communication
Improving classroom teaching
Navigation
MOOC

ABSTRACT

Although massive open online courses or MOOCs have been successful in attracting a large number of learners, they have not been equally successful in retaining the learners to the point of course completion. One critical point of failure in many courses, especially those that use discussion forums as a means of collaborative learning, is the large number of messages exchanged on the forums. The extensive exchange of messages often creates chaos from the instructors' perspective and several questions remain unanswered. Lack of attention and response to urgent messages – those that are critical from the learners' perspective to move forward – becomes a major challenge in this environment. This paper proposes a model to identify "urgent" posts that need immediate attention from instructors. In our analysis, we investigate different feature sets and different data mining techniques, and report the best set of features and classification techniques for addressing the problem of identifying messages that need urgent attention. The results demonstrate the ability to use a limited number of linguistic features with select metadata to build a moderate to substantially reliable classification model that can identify urgent posts in MOOC forums regardless of the course content. The work has potential application across a range of platforms that provide large scale courses and can help instructors efficiently navigate the discussion forums and prioritize the responses so that timely intervention can support learning and may reduce dropout rates.

1. Introduction

Since their inception in 2008, Massive Online Open Courses or MOOCs have witnessed remarkable growth in number of participants, courses offered, and availability of different channels that offer MOOCs. The goal behind modern MOOCs is to provide global access to open online resources on a large scale (Liyanagunawardena, Adams, & Williams, 2013). After nearly a decade, more than 58 million users are estimated to have registered for at least one MOOC and more than 700 universities are offering thousands of courses that reside on different platforms such as Coursera and edX (Shah, 2016). The most recent trend in MOOC development is offerings that go beyond individual courses and offered massive open online degrees e.g. Georgia Tech's in collaboration with Udacity (Master of Science in Computer Science) and edX (Master of Science in Analytics) (Maderer, 2017).

Within MOOCs, discussion forums provide an important platform for online learners to interact with each other and with the instructor. The use of discussion forums reduces a sense of isolation among learners and allows them to share knowledge and concerns. However, discussion forums with large number of postings (and participants) in an online learning environment decrease the level of interactivity among learners (Kim, 2013). Furthermore, with hundreds of posts in an online class, it is difficult for

* Corresponding author. 4509 Nguyen Engineering Building, George Mason University, Fairfax, VA 22030, USA.
E-mail addresses: oalmatra@masonlive.gmu.edu (O. Almatrafi), johri@gmu.edu (A. Johri), rangwala@cs.gmu.edu (H. Rangwala).

https://doi.org/10.1016/j.compedu.2017.11.002
Received 14 May 2017; Received in revised form 1 November 2017; Accepted 8 November 2017
Available online 12 November 2017

instructors to review all the posts questions and comments. As a result, there has been a call for an alternative representation for the written data in discussion forums so that instructors can effortlessly have a comprehensive overview of the information embedded in the discussion (Dringus & Ellis, 2005) and be informed when new posts of interest are published (Lin, Hsieh, & Chuang, 2009).

In this paper we present a study whose aim is to develop a supervised learning model that can automatically identify urgent posts in MOOC discussion forums regardless of the course topic or content. Urgent posts are categorized as those posts that call for immediate attention from the instructors. Although, based on socio-constructive theories, the role of instructors in online learning entails more than just responding to critical posts (for instance, increasing engagement, promoting deep learning, and creating a sense of community) (Woo & Reeves, 2007), our focus in this study is on identifying critical issues as represented in MOOC discussion forums. We believe that an efficient mechanism for monitoring and responding to urgent posts will help instructors prioritize their responses and better manage the large volume of posts. It will also help free up instructor time and attention to engage in more community building and scaffolding activities. In addition, the model may lead to less confusion and a higher completion rate as the instructors can intervene in a timely manner. Research has shown a correlation between confusion and dropout (Yang, Wen, Howley, Kraut, & Rose, 2015) and studies also show that lack of responsiveness in MOOC forums could be a factor contributing to learner dropout (Hone & El Said, 2016; Wang, Yang, Wen, Koedinger, & Rosé, 2015).

In the remainder of this paper, we will briefly introduce related work (Section 2). Then, we will explain the current study, the data, and methods in Section 3. After that, we will report the results of the analysis in Section 4 and discuss them in Section 5. Lastly, in Sections 6 and 7, we will draw attention to some limitations and suggest future work.

2. Literature review

2.1. MOOCs and discussion forums

In online learning, it is important to facilitate interaction among learners and between learners and instructors to ensure a high quality learning experience and discussion forums are one of the tools used by instructors for this purpose (Richardson et al., 2015). In addition to cognitive benefits of engaging in dialogue, MOOC learners use discussion forums to network, report problems, express opinions, seek clarification on materials, and form teams to collaborate and gain a better understanding of the material (Wise, Cui, Jin, & Vytasek, 2017). It has been reported in a study that examined MOOCs discussion forums that the average number of newly created posts in three MOOCs courses with 1146, 771, and 24,963 active participants are 96, 152, and 510 posts per day, respectively (Wen, Yang, & Rosé, 2014b). In a larger seven-week MOOC with 50,000 + enrollments the reported number of posts exceeded 50,000 (Wong, Pursel, Divinsky, & Jansen, 2015). This amount impedes instructors and learners to effectively navigate the discussion forums to find messages relevant to their purpose. In addition, there is a high ratio of instructors to learners, which suggests that instructors should utilize their capacity wisely and be selective in the intervention (Chaturvedi, Goldwasser, & Daumé, 2014) given that a large number of discussion forum messages are not content related (Wise et al., 2017). In a study that interviewed MOOC instructors, instructors emphasized the need for better ways to navigate MOOC discussion forums as they quickly become overwhelming (Hollands & Tirthali, 2014, pp. 1–208). One instructor, in that study, recommended the use of natural language processing to organize the discussion forums as a step to face the problem of complexity at a broader level.

Different approaches have been utilized to address the problem of disorder in discussion forums in MOOCs including classification, and recommendation. Each approach contributes to the literature differently. Recommendation suggests potential posts of interest based on users past behavior, or suggests a resource (e.g. a clip of a video lecture) that best matches the confusion expressed in the post (Yang, Piergallini, Howley, & Rose, 2014; Yang et al., 2015). A recent study proposed an approach to route questions to potential participants who can answer the question based on their willingness and knowledge expertise. The authors also observed that some MOOC questions cannot be answered by other leaners and require instructors' response (Macina, Srba, Williams, & Bielikova, 2017). Classification, on the other hand, was used to identify specific dimensions of posts based on pre-defined categories. Classification has been used in the literature to study MOOC discussion forums and identify content-related posts, posts expressing confusion, and sentiment (Agrawal, Venkatraman, Leonard, & Paepcke, 2015; Wen et al., 2014b; Wise et al., 2017).

Chaturvedi et al. (2014) first introduced the problem of predicting instructors' interventions in MOOC discussion forums. A similar study in purpose was conducted by Chandrasekaran, Kan, Tan, and Ragupathi (2015). Both proposed models that predict the intervention for Coursera MOOCs at a thread level, which conceal the identification of new posts to a thread. Another issue is that the ground truth for intervention decisions were based on the threads that instructors have intervened to. The problem is that instructors adopt different strategies of teaching and intervention. Hence, models trained on subjective interventions are biased and problematic to generalize at a global level.

2.2. Methodological approaches

There are several theoretical frameworks to categorize and label MOOC posts. Among the earliest was a framework that categorizes MOOC posts into two dimensions – the topic of the post and the poster role (Stump, Deboer, Whittinghill, & Breslow, 2013). The topic of the posts is further divided into different sub-categories: content, social/affective, course website/technology, course structure/policies, etc. and the poster role includes: help-seeker, help-giver, or other. Another classification model classified posts according to six dimensions: question, answer, opinion, confusion, sentiment, and urgency (Agrawal et al., 2015); each post takes a value in each dimension. Question, answer and opinion are binary variables, while the other dimensions can take a discrete value in the range from 1 to 7. In our analysis, we used a dataset that was coded based on the latter model. It is worth mentioning that urgent