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
0360-1315/

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

posts can express urgency regarding content, course website or course policies.

A significant contribution to the effort of bringing order to the chaos in MOOC discussion is the classification of posts into content vs. non-content related posts (Cui & Wise, 2015). Wise et al. (2017) applied logistic regression to bag-of-words features (unigram and bigram) and validate the model on the same subject domain, and in the following research they tested their model on different disciplines. On average the model showed an accuracy of 77%, Kappa value of 0.57, recall of 66% and precision of 65%. This model is beneficial for many applications, but from an instructor's point of view, identifying content-related posts doesn't help in identifying posts that require immediate response.

Another related work is the identification of confusion in MOOC posts (Agrawal et al., 2015; Yang et al., 2015). Agrawal and his co-authors first identified posts as confusing and followed that with a recommendation regarding the start time of a video clip from an online lecture that specifically addresses the confusion. The classification model was built using features that include: bag-of-words, post metadata, and the prediction of the post being a question, an answer, an opinion, a sentiment, and an urgent post. The model was trained using standard logistic regression. It performed as follows: Kappa = 0.62, F1 = 0.70 for the humanities domain, and Kappa = 0.36, F1 = 0.627 for the education domain. Although confusion and urgency are correlated in two out of three tested domains, they imply different dimensions; confusion measures the level of confusion the student exhibits in the post (e.g. “Which article were you thinking of using?”), while urgency measures the level of criticality of the post to get instructors’ attention and intervention (e.g. “Week 2 is loading but I cannot access Chapter 3 … says “its restricted from the server".is anyone else having issues with this?”).

Last but not least, there was an effort to develop a general classification model that can classify posts according to three dimensions: sentiment, urgency, and confusion across different domains (Bakharia, 2016). The work is still in its infancy and fails to generalize the classification of urgent posts. The classification model may work well when it is tested on a similar course or within a domain but does not generalize well when it is applied to dissimilar courses and domains. One of the principal conclusions is a recommendation to investigate the types of algorithms that can be adopted within an educational context. Hence, we aim to build an urgency-detection model that can identify urgent posts in MOOC discussion forums at a global level using linguistic features.

Our analysis is different from prior works in the following ways: (i) we focus on urgency identification, which is important for instructors to prioritize their responses; (ii) we compare different sets of features and classifier methods to report which one best predicts the urgency of the posts; and (iii) we validate the model on different samples of the dataset to report its generalizability within the current platform.

3. Current study

The goal is to examine the possibility of building a generalizable and robust model that can classify posts based on urgency using linguistic features. To achieve the goal, several research questions need to be addressed.

RQ1: Can linguistic features such as term frequency and features extracted from a linguistic tool along with some metadata identify reliably urgent posts in MOOC forums?

- RQ1.a: What are the most predictive linguistic features?
- RQ1.b: How well does each linguistic feature-extraction method perform in the current study?

RQ2: Which classification technique is the best to identify urgent posts reliably?
RQ3: How well does the model identify urgent posts?

- RQ3.a: using all the data
- RQ3.b: omitting some courses
- RQ3.c: omitting a subject domain

3.1. Data source

We used The Stanford MOOCPosts dataset (Agrawal & Paepcke, 2014) in our analysis. This dataset contains a large number of posts – 29,604 collected from eleven MOOC courses. Those courses were selected from three domains equally: Humanities, Medicine, and Education. Each post was coded by three human coders generating the gold sets (more information regarding score computation can be found in the dataset website (Agrawal & Paepcke, 2014)). Each post was coded by humans on various dimensions as mentioned in the literature review section, hence the data provides a ground truth for the proposed model. In our analysis, the full dataset includes 29,584 posts after we excluded posts that have no meaningful information on their own such as ‘+1’, or ‘75%’. It would have meaning in the thread but the unit of analysis in this study is the post.

In the urgency dimension, coders ranked the urgency of the post on a scale of 1–7. A score of one means the post is not urgent and no response is required from the instructors, while 7 means it is extremely urgent and requires immediate attention from the instructors. Table 1 exhibits examples of posts and their urgency scores. We also examined the dataset and found that among urgent posts about 66% express opinion, and have sentiment score ≥ 4, and approximately 73% coded as question while 7% as answer. Most of the urgent posts express confusion (approx. 96%).

In our analysis, our goal is to assess whether the post is urgent or not (binary classification). We considered urgent posts to be the ones scoring 4 or above, otherwise the post in not urgent. With this binary categorization, we found that urgent posts constitute about

Table 1
Examples of MOOC discussion forums posts and its human-coded urgency score.

Post Example	Urgency Score
Professor, This was an awesome analysis. Now probability started making sense to me:)	1
How long would it take to grade the peer assessments?	4
Answers to QUIZ: PRICE CEILINGS AND FLOORS are wrong. The explanation disagrees with the marked answers. Please check!	7

20% of all posts (starting posts and comments) across all domains. Since the data is imbalanced, some considerations such as splitting the data, training the model, and choosing the evaluation metrics were considered.

3.2. Methods

3.2.1. Feature extraction methods

3.2.1.1. Linguistic Inquiry and Word Count (LIWC). Linguistic Inquiry and Word Count (LIWC 2015) is an application that analyzes text linguistically from different dimensions. It measures the social and psychological meaning of words such as emotions, attention, and cognitive thinking expressed in the text. LIWC has been utilized in content analysis research in educational contexts. For instance, Wong et al. (2015) used it to analyze MOOC discussion forums and to examine the use of some linguistic features such as pronouns and affective processes over the course time. Wen, Yang, and Rosé (2014a) used LIWC's cognitive indices in conjunction with other linguistic features, to measure learners' cognitive engagement in MOOCs. In our analysis, LIWC was used to derive linguistic features for MOOC posts. Each post was projected into 94 features. We used the whole set and then selected a subset (the most important) without deteriorating the classification performance.

3.2.1.2. Metadata. We also incorporated three of the posts' metadata features, which are: 'up_count', 'reads', 'post_type'. Up_count and reads features show the number of reads and up votes for the post. The post_type is essentially the type of post, which is either a comment or a starting post of the thread.

3.2.1.3. Term Frequency (TF). The bag-of-words model was used to represent the text in the posts. More specifically, we used unigram term frequency, which means each word in a post was assigned a weight based on the number of occurrences of the word in the post. Stop words and pronouns were not removed from the text. Their inclusion improved content-related classification for MOOC discussion forums (Wise et al., 2017). We assume they may be useful in our problem.

3.2.2. Classification methods

In this study, we evaluated five different classification approaches, namely, Naïve Bayes, Support vector machines, Random forests, AdaBoost (decision trees as base estimators) and Logistic regression. In general, naïve Bayes has the lowest performance of all; therefore, we omit it. LinearSVC had almost similar results to logistic regression, so we omit linearSVC for brevity of presentation.

We used default parameters. However, to address imbalance in the distribution of urgent posts versus non-urgent posts we penalized the misclassification loss to be inversely proportional to the class frequencies in dataset.

Since we have imbalanced data, F1-weighted and Cohen's Kappa are used as metrics for model evaluation. F1-weighted computes the harmonic mean between recall and precision for both classes. Cohen's Kappa is used to measure the inter-rater reliability. It computes the accuracy of the model considering agreement by chance.

3.2.3. Sampling groups

We have tested the model on different samples of the dataset to assess its applicability independent of the domain or course the posts belong to. Three groups of data were created for training and testing (Table 2):

- *Group A* (All data in): In this group, all the data "posts" was utilized. We divided the data into two datasets: training and testing datasets. The training dataset consists of 66% of the overall data. The training set and testing set are split using stratified sampling to ensure the original ratio of "Urgent" variable is preserved.
- *Group C* (some Courses out): We split the data into training and testing based on the course name. For testing, we chose some courses from Humanities and Medicine, namely, Stat Learning (Winter 2014), Statistics in Medicine, and Managing Emergencies: What Every Doctor Must Know? Education domain, however, has only one course, thus we reserved 33% of stratified education

Table 2
The training and testing split in different samples of the dataset and their urgency distribution.

Sample	Training #posts (%urgent)	Testing #posts (%urgent)
Group A (All in)	19,821 (22%)	9763 (22%)
Group C (3 courses & 33% of Education course for testing)	19688 (21%)	9883 (23%)
Group D (Humanities for testing)	19871 (20%)	9711 (26%)

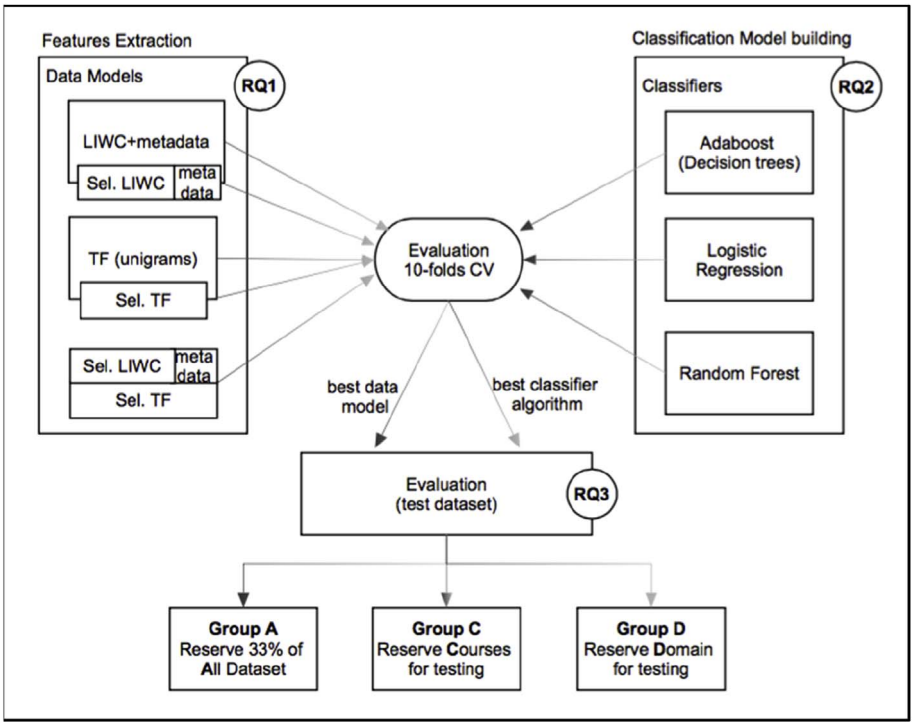

Fig. 1. The flowchart illustrates the process followed in this study.

data to testing. The choice of courses to be in the testing dataset was a balance between removing courses from the domain and presevering the percentage of training and testing data to be close to 66 and 33% respectively.

- *Group D* (a Domain out): In this group, we reserved a domain for testing. We trained our classifier on Medicine and Education domains and evaluated the result on Humanities domain. The choice of Humanities to be held out for testing is arbitrary.

The flowchart in Fig. 1 illustrates the stages and processes followed in this study, starting from feature extraction, choice of classification model-building algorithm, and concluding with validating the chosen data model and the classifier model on different datasets.

4. Results

In this section we presents results organized according to the research questions posed earlier.

The first research question is regarding the ability for linguistic features derived from LIWC and some metadata to identify urgent posts. To answer this question, we need to know:

RQ1 (a &b): How well does each linguistic feature-extraction method (data model) perform in the current study? What are the most predictive linguistic features?

As earlier mentioned in the feature extraction methods section, two main ways were used to extract features from the posts plus posts metadata. Metadata has been added to LIWC features to constitute LIWC + metadata model. The other model used unigram TF model. Table 3 summarizes the performance using 10-fold cross-validation when the best performing classification techniques was applied on Group A-training dataset. We ran the analysis on the full set of features and on a subset of selected ones. The best features were selected using Chi-squared test. Each feature is scored on its importance of predicting the outcome 'Urgency'. In this study, p-

Table 3
Comparing the performance of different data representation models and features selection.

Data Model (Classifier algorithm)	Metrics	All Features	Selected Features (p-value < 0.001)
LWIC + Metadata (AdaBoost)	# Features	97	80
	F1-weighted	0.87	0.87
	Kappa	0.59	0.59
TF (Logistic Regression)	# Features	28,074	137
	F1-weighted	0.84	0.80
	Kappa	0.57	0.46

Table 4
The most important features in each data model (sorted by importance).

Data Model	Most important Features
LIWC + Metadata	'Clout','posemo','reads','Exclam','affect','Tone', 'Authentic', 'social','number','they','assent','informal','QMark','you', 'tentat'
TF	'please', 'math', 'students', 'peer', 'submitted', 'essay', 'quiz', 'submit', 'any', 'mistakes', 'advance', 'anyone', 'does', 'grading', 'question'

value was used as a threshold to select the important ones. The classifier performance and the number of features for each model is presented in Table 3.

All three metadata features made it to the selected features when they were combined with LIWC features. Table 4 shows the fifteen most important features sorted by their importance in the two data models LIWC + Metadata and TF.

The TF seems to give a good indication of the topics in urgent posts. Out of the 137 words nearly 10% were course-specific such as 'math', 'mean', and 'talk'. Other features were very broad and can be generally categorized into the following groups presented in Table 5.

RQ2: Which classification technique is the best to identify urgent posts reliably?

For this question, we employed Group A to evaluate different classifiers on different data models. Group-A does not separate posts based on course or domain. Around Sixty-six percent of all MOOC posts in the dataset are in the training dataset. Fig. 2 shows the performance of the best three tested classifiers. Ten-fold cross-validation was utilized to obtain the performance metrics, which is weighted-F1 and Cohen's Kappa. The best model was the one trained using AdaBoost and operated on LWIC + TF features, with the following metrics: F1-weighted (0.88) and Kappa (0.64). Adaboost tends to work better with imbalanced data especially if time is not a big issue. This is because Adaboost gives more emphasis to misclassified instances in the subsequent iteration.

In general, AdaBoost was the best classifier algorithm for LIWC + metadata, and LIWC + TF. In contrast, logistic regression did better when operating on TF data features alone. Random Forests takes the second place after AdaBoost classifier in especially when using LIWC + metadata or LIWC + TF. Using LIWC + metadata (80 features) is most promising to identify urgent posts. Comparing LIWC + metadata and TF, LIWC + metadata produces more useful features for the classification. Combining both features always results in better classification models.

RQ3: How well does the model identify urgent posts? (Using all the data, omitting some courses, and omitting a domain)

After identifying the best data model and classifier algorithm, we assessed the validity of the feature extraction approach and the classification technique in classifying urgent posts when the data trained is missing some courses (Group C) or a domain (Group D). Therefore, we evaluated the performance of our approach on a testing unseen dataset (Table 6). As seen in the table, Group A and Group C have almost similar performance (weighted-F1 = 0.88, F1 ≥ 0.70 and Kappa ≥ 0.63). Group D is not as good as the other groups, yet has a moderate performance.

5. Discussion

The aim of this research is to facilitate instructors' role in MOOCs, more specifically, assist them in navigating students' posts in MOOC discussion forums in a more efficient and effective way. The goal of the study was to examine the possibility of building a general, reliable model that can identify urgent posts in MOOC discussion forums regardless of the course domain. Different feature extraction methods and data mining techniques were examined to build a reliable model. The work was inspired by (Agrawal et al., 2015; Wise et al., 2017) who used classification techniques to identify confusion and content-related posts respectively. This research extends the prior work to identify urgent posts that are important from the instructors' perspective to provide timely intervention. Moreover, in order to build a general model, it has been evaluated on a holdout set of unseen courses (group C) and a domain (group D). In this section, we discuss the results presented in the previous section.

RQ1: Can linguistic features such as term frequency and features extracted from LIWC along with some metadata reliably identify urgent posts in MOOC discussions?

Our analysis shows that LIWC alone was able to perform reasonably well and when the features derived from LIWC joined with selected TF, the combined model performed best. The unigram features (TF) emphasized the results in (Cui & Wise, 2015; Wise et al.,

Table 5
Examples of the frequent terms that were used to train the model grouped into several categories.

Category	Example of words
Polite/message	please, pls, plz, hi, kindly, dear, etc.
Course-related	essay, quiz*, submi*, certificate, assignment*, homework, module, video, validation, lectures, questions, etc.
Grading	feedback, grad*, mark*, review*, peer, etc.
Warning	mistakes, incorrect, advance, error, completed, missing, invalid, wrong, unable, accidentally, etc.
Technical	access, technical, session, progress, button, saved, upload, download
Thoughts	wonder*, think*, confus*, what, why, etc.
Pronouns	any*, my, me, them, and your
Conjunctions	but, or, still, etc.

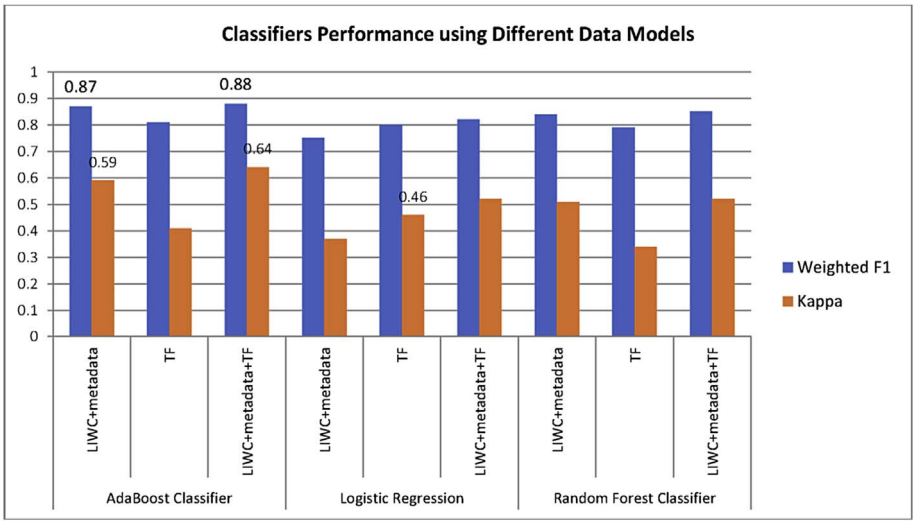

Fig. 2. Classifiers Performance operating on different data models.

Table 6
The performance metrics on holdout dataset (AdaBoost classifier operating on LIWC + TF data model).

	Not Urgent			Urgent			Weighted- F1	Kappa
	P	R	F1	P	R	F1		
Group A	0.91	0.95	0.93	0.77	0.65	0.70	0.88	0.63
Group C	0.90	0.95	0.92	0.80	0.65	0.72	0.88	0.64
Group D	0.87	0.95	0.91	0.80	0.57	0.67	0.85	0.58

*P: precision, R: recall.

2017) that top features were terms related to the process of learning, question words, and terms that connect ideas. Additionally, in our case, thoughts, warning words, and some general course and technical words were among the top features. Also, pronouns were in the top features, which played a role in identifying cognitive engagement (Wong et al., 2015).

In building the model, all posts were included regardless of the post type (starting post or comment). In fact, post type was one of the metadata examined in the model; and it made it to the top features in the LIWC + metadata model along with the other metadata features (up_votes and reads). As shown in Table 3, in the experiment, a subset of LIWC + metadata features performed as well as including all the features.

The selected most predictive variables derived from LIWC shed a light on the linguistic and psychological features learners used to express urgency. Psychologically, many features fall into one of the three big groups: confidence, tone and authenticity. The use of language also contributes to the prediction, such as the use of pronouns and punctuations. Urgent posts have tendency to include more question marks, numbers, and tentative and authenticity language; at the same time, features such as clout, tone, social and others presented in Table 4 are negatively correlated to urgency. In sum, using linguistic features derived from LIWC + metadata alone showed moderate reliability.

RQ2: Which classification technique is the best to reliably identify urgency of posts?

The best performing algorithm for the problem and dataset used in this study is AdaBoost (with decision trees as weak learners). This is true when the features are derived from LIWC or unigrams joined with LIWC. Random forest comes second. On the other hand, logistic regression surpasses other algorithms when the features are unigrams (TF) only. In section 4, we demonstrate the performance for Group A. Nonetheless, similar results regarding the best classifier were achieved when trained and tested on other groups (Group C and Group D).

RQ3: How well does the model identify urgent posts?

The results are promising; the model is moderately to substantially reliable in identifying urgent posts (lowest Kappa: 0.58; highest: 0.64) when tested on unseen courses and a domain. According to Landis and Koch (1977), Cohen's Kappa scores between (0.41–0.60) are considered moderate and (0.61–0.80) substantial. This performance was achieved with at most 262 linguistic features combined from LIWC, metadata and unigrams (term frequency), which is another advantage of the model. The availability of a simple model with few features speeds up the computations in the classification process.

6. Limitations

Our model utilizes fewer than 300 linguistic features to identify urgent posts across courses and domains. The simplicity of the model is an advantage and demonstrates the ability to achieve the intended goal. However, there is still some room for performance improvement, most importantly, increasing the recall for urgent posts without deteriorating the overall performance. For instance, tuning algorithms' parameters, or using another data mining techniques such as anomaly detection.

A limitation with the current study is the limited data available. Incorporating clickstream data such as repeated clicks on the same item may assist with better understanding issues faced by learners, but it is not clear whether those issues represent urgency or confusion. Based on the way confusion and urgency operationalized, we expect clickstream data would be helpful to identify confusion more than urgency. Adding metadata, however, about the course information and forum structure will most likely improve the classification (Chaturvedi et al., 2014). For instance, forum type, which refers to the forum title that posts originated from (ex. lectures, exams, etc.), improved the classification (Chandrasekaran et al., 2015). Furthermore, the time the post was created could be a critical feature. A couple of days before some course activities deadline may show an improvement in the model performance, yet need to be inspected.

7. Future work

With respect to generalizability, the features and the model proposed have reliably identified urgent posts for Stanford MOOC discussion forums data. Future work can investigate the external validity of the model on other MOOC discussion forums platforms.

One natural next step is to investigate the best approach to demonstrate the results of the model to instructors to allow them respond efficiently to urgent posts. Integrating successful models into MOOCs' platform may take different tactics, for example, using visualization, or organizing and prioritizing the posts in a list. In fact, computing a probabilistic score for the level of urgency could give a finer grain into the criticality of the post, which we will examine next.

Another possible direction for future work is to dig deeper into the origin causes of urgent posts by investigating the topic vented in urgent posts according to the categories mentioned in (Stump et al., 2013). Is the topic expressed in the post is content-, logistic- or technical-related? Consequently, future research can attempt to mitigate the causes or automate the intervention when appropriate.

8. Conclusion

This paper demonstrates the ability to use a limited number of linguistic features with few metadata to build a moderately to substantially reliable classification model that can identify urgent posts in Stanford MOOC discussion forums regardless of the course or domain the posts belong to. In addition, we examined different linguist-features groups: linguistic features derived from LIWC, and unigram bag-of-words along with few metadata. We also surveyed several data mining techniques, specifically, Support vector machine, Naïve Bayes, Logistic regression, Random forests, and AdaBoost. We evaluated our model in 10-fold cross-validation and against holdout testing set to mitigate over fitting and test the validity of the introduced model across courses and domains.

The work has a potential impact in maximizing instructors' efficiency in monitoring large online discussion forums. The ultimate goal is to provide instructors with real time identification of urgent posts so that they can respond on time. As a result, it may lead to less confusion and higher completion rates as the instructors intervene in a timely manner. Another advantage of the automation of classification in MOOC forums is that it can help optimize human resources to focus on improving the learning experience.

References

Agrawal, A., & Paepcke, A. (2014). *The Stanford MOOCPosts data set.* Retrieved July 10, 2017, from http://datastage.stanford.edu/StanfordMoocPosts/.

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). YouEDU: Addressing confusion in MOOC discussion forums by recommending instructional video clips. *Proceedings of the 8th international conference on educational data mining* (pp. 297–304). .

Bakharia, A. (2016). Towards cross-domain MOOC forum post classification. *Proceedings of the third (2016) ACM conference on learning @ scale - L@S '16* (pp. 253–256). . http://doi.org/10.1145/2876034.2893427.

Chandrasekaran, M. K., Kan, M.-Y., Tan, B. C. Y., & Ragupathi, K. (2015). Learning instructor intervention from MOOC Forums: Early results and issues. *Proceedings of the 8th international conference on educational data mining* (pp. 218–225). . Retrieved from http://arxiv.org/abs/1504.07206.

Chaturvedi, S., Goldwasser, D., & Daumé, H. (2014). Predicting instructor's intervention in MOOC forums. *52nd annual meeting of the association for computational linguistics, ACL 2014-proceedings of the conference: Vol. 1*, (pp. 1501–1511). . Baltimore, Maryland, USA. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84906924486&partnerID=tZOtx3y1.

Cui, Y., & Wise, A. F. (2015). Identifying content-related threads in MOOC discussion forums. *Learning at Scale,* 299–303. http://doi.org/10.1145/2724660.2728679.

Dringus, L. P., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers and Education, 45*(1), 141–160. http://doi.org/10.1016/j.compedu.2004.05.003.

Hollands, F., & Tirthali, D. (2014). *MOOCs: Expectations and reality.* Center for Benefit-Cost Studies of Education, Teachers College, Columbia University, (May)http://doi.org/10.1109/EDOC.2006.60.

Hone, K. S., & El Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study. *Computers and Education, 98*, 157–168. http://doi.org/10.1016/j.compedu.2016.03.016.

Kim, J. (2013). Influence of group size on students' participation in online discussion forums. *Computers & Education, 62*, 123–129. http://doi.org/10.1016/j.compedu.2012.10.025.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159. http://doi.org/10.2307/2529310.

Lin, F. R., Hsieh, L. S., & Chuang, F. T. (2009). Discovering genres of online discussion threads via text mining. *Computers and Education, 52*(2), 481–495. http://doi.org/10.1016/j.compedu.2008.10.005.

Liyanagunawardena, T. R., Adams, A. A., & Williams, S. A. (2013). MOOCs: A systematic study of the published literature 2008-2012. *International Review of Research in Open and Distance Learning, 14*(3), 202–227. http://doi.org/10.3329/bjms.v12i4.16658.

Macina, J., Srba, I., Williams, J. J., & Bielikova, M. (2017). Educational question routing in online student communities. *Proceedings of the eleventh ACM conference on recommender systems* (pp. 47–55). Como, Italy: ACM. http://doi.org/10.1145/3109859.3109886.

Maderer, J. (2017). *Georgia Tech creates first online master of science in Analytics degree for less than $10,000*. Retrieved April 20, 2017, from https://pe.gatech.edu/news/press-release/01072017.

Richardson, J. C., Koehler, A.a, Besser, E. D., Caskurlu, S., Lim, J., & Mueller, C. M. (2015). Conceptualizing and investigating instructor presence in online learning environments. *International Review of Research in Open & Distance Learning, 16*(3), 256–297. http://doi.org/10.1080/01587919.2015.1055920.

Shah, D. (2016). *By the numbers: MOOCS in 2016*. Retrieved April 20, 2017, from https://www.class-central.com/report/moocs-2015-stats/.

Stump, G. S., Deboer, J., Whittinghill, J., & Breslow, L. (2013). Development of a framework to classify MOOC discussion forum Posts: Methodology and challenges. *NIPS workshop on data driven education, (December)* (pp. 1–20). .

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains. *Proceedings of the 8th international conference on educational data mining* (pp. 226–233). . Retrieved from http://www.educationaldatamining.org/EDM2015/proceedings/full226-233.pdf.

Wen, M., Yang, D., & Rosé, C. P. (2014a). Linguistic reflections of student engagement in massive open online courses. *Proceedings of the 8th international conference on weblogs and social media ICWSM*.

Wen, M., Yang, D., & Rosé, C. P. (2014b). Sentiment analysis in MOOC discussion Forums: What does it tell us? *Proceedings of the 7th international conference on educational data mining* (pp. 130–137). . Retrieved from http://www.cs.cmu.edu/~mwen/papers/edm2014-camera-ready.pdf.

Wise, A. F., Cui, Y., Jin, W. Q., & Vytasek, J. (2017). Mining for gold: Identifying content-related MOOC discussion threads across domains through linguistic modeling. *Internet and Higher Education, 32*, 11–28. http://doi.org/10.1016/j.iheduc.2016.08.001.

Wong, J. S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015). Analyzing MOOC discussion forum messages to identify cognitive learning information exchanges. *Proceedings of the association for information science and technology: Vol. 52*, (pp. 1–10). . http://doi.org/10.1002/pra2.2015.145052010023.

Woo, Y., & Reeves, T. C. (2007). Meaningful interaction in web-based learning: A social constructivist interpretation. *Internet and Higher Education, 10*(1), 15–25. http://doi.org/10.1016/j.iheduc.2006.10.005.

Yang, D., Piergallini, M., Howley, I., & Rose, C. (2014). Forum thread recommendation for massive open online courses. *Educational Data Mining, 2014*, 257–260.

Yang, D., Wen, M., Howley, I., Kraut, R., & Rose, C. (2015). Exploring the effect of confusion in discussion forums of massive open online courses. *Proceedings of the second (2015) ACM conference on learning @ scale - L@S '15* (pp. 121–130). . http://doi.org/10.1145/2724660.2724677.