Towards Improving Students' Forum Posts Categorization in MOOCs and Impact on Performance Prediction

Fatima Harrak
Sorbonne Université
CNRS, Laboratoire
d'Informatique de Paris 6
F-75005, Paris, France
fatima.harrak@lip6.fr

François Bouchet
Sorbonne Université
CNRS, Laboratoire
d'Informatique de Paris 6
F-75005, Paris, France
françois.bouchet@lip6.fr

Vanda Luengo
Sorbonne Université
CNRS, Laboratoire
d'Informatique de Paris 6
F-75005, Paris, France
vanda.luengo@lip6.fr

Rémi Bachelet
Centrale Lille
University of Lille
France
remi.bachelet@ec-lille.fr

L@S '19, June 24–25, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6804-9/19/06.
DOI: https://doi.org/10.1145/3330430.3333661

Abstract
Going beyond mere forum posts categorization is key to understand why some students struggle and eventually fail in MOOCs. We propose here an extension of a coding scheme and present the design of the associated automatic annotation tools to tag students' questions in their forum posts. Working of four sessions of the same MOOC, we cluster students' questions and show how the obtained clusters are consistent across all sessions and can be sometimes correlated with students' success in the MOOC. Moreover, it helps us better understand the nature of questions asked by successful vs. unsuccessful students.

Author Keywords
Student's question; discussion forum; coding scheme; clustering; student's performance; MOOC

ACM Classification Keywords
I.5.3 [Pattern recognition]: Clustering; K.3.1 [Computers and Education]: Computer Uses in Education

Introduction
Students' questions can be used to improve their learning experience and help teachers better understand their thinking. In MOOCs, discussion forums are a key feature and it has been shown that lack of social interaction as a valuable form of learning is one of the main concern [4]. Although

	N1	N2	N3
GDP5	17579	7655	2087
GDP6	23315	10597	4717
GDP7	19392	12224	3504
GDP8	24603	14072	4760

Table 1: Descriptive statistics of the 4 MOOC sessions considered (registration, messages and success)

N1 = Students registered
N2 = Nb of posts
N3 = Nb of unique posters

Kappa between the automatic annotator and the manual annotation
$\kappa_q = 0.91$
Question classification (DT)
$\kappa_c = 0.66$
Course classification (SVM)
$\kappa_0 = 0.37$
dimension 0 classification (GBT)
$\kappa_1 = 0.68$
dimension 1 classification (GBT)
$\kappa_2 = 0.39$
dimension 2 classification (GBT)
$\kappa_3 = 0.56$
dimension 3 classification (GLM)
$\kappa_4 = 0.48$
dimension 4 classification (DT)

several works have tried to show the impact of categorizing students' posts [5], whether they are content-related [6, 2], urgent [1], they rarely look into the detailed content of the posts. We hypothesize that analyzing more finely the content of MOOC posts would help in particular to predict students' success.

We want to address two research questions: (RQ1) Can we reliably annotate questions extracted from MOOC forum posts according to a fine-grained multi-level coding scheme? And (RQ2) is there a consistent relationship between students' questions and their performance in the MOOC? We address RQ1 by extending an existing coding scheme and developing a coding tool using several classifiers in cascade to annotate sentences extracted from students' posts on a MOOC. Then we address RQ2 by using clustering to investigate whether the type of questions asked by students relate to their success.

Research Context

We consider log and forum data from four different sessions of the same French MOOC on project management called GDP (French acronym for project management) held in 2015 and 2016 (sessions 5 to 8). The forum works in a typical manner, organized around threads created by the pedagogical team to answer to technical or administrative issues, about homework or course content, among others. Table 1 provides some basic statistics on the forum usage and number of students registered. For each session we extracted the students' posts in course related topics, the final grade (out of 100) and whether students were successful or not (grade superior to 50%).

Question Coding Scheme

Coding scheme design

We considered a sample of 500 messages from the 4 sessions randomly divided into 3 sub-samples (200/100/200) to apply 3 successive categorization steps to define a coding scheme as proposed by Harrak et al. in another context of study [3]. The raw corpus contains messages posted in the discussion forum and can be very unstructured and noisy (*e.g.* a message can contain several questions, opinions, answers to issues not course related, *etc.*). We excluded from this analysis: (1) messages coming from the instructors, (2) messages that are a reply to other ones (*e.g.* not the root messages), and (3) the threads which are explicitly not course related (e.g. a thread on technical issues).

The messages were segmented into several questions (using NLTK) and annotated according to their content. The course-based questions were annotated given the coding scheme from [3]. This coding scheme (cf. Table 2) consists in tagging each question according to 4 independent dimensions: a main mandatory one (dimension 1), and 3 optional ones (dimensions 2 to 4). For instance, a question could be a request to re-explain the way something work by providing another example (tagged as "Ree" on dimension 1, "Exa" on dimension 2, "Man" on dimension 3, and nothing (0) on dimension 4, *i.e.* [Ree,Exp,Man,0]). The non-course related questions were then annotated according to dimension 0 (*cf.* Table 3). Two human annotators used as a unique reference the coding scheme introduced in Tables 2 and 3 to annotate each question. Their agreement was evaluated using Cohen's Kappa (average value: 0.71).

Automatic annotation

To annotate the entire corpus of messages, we performed the classical preprocessing steps on the training sample of 1307 segments (500 messages) manually annotated: to-

Dim1	Question type
Ree	**Re-e**xplain / redefine
Dee	**Dee**pen a concept
Ver	Validation / **ver**ification
Dim2	**Explanation modality / Quest. subject**
Exp	**Ex**am**p**le
Sch	**Sch**ema
Cor	**Cor**rection
Dim3	**Explanation type**
Def	**Def**inition
Man	**Man**ner (how?)
Rea	**Rea**son (why?)
Rol	**Rol**es (utility?)
Lin	**Lin**k between concepts
Dim4	**Verification type**
Mis	**Mis**take / contradiction
Kno	**Kno**wledge in course
Exp	**Exp**ected knowledge in assessment

Table 2: Coding scheme used to tag course-based students' questions (adjusted from (Harrak et al. 2018))

kenization, stemming, punctuation removal (except for '?') and stopwords (non-meaningful words). We then counted the occurences of all the unigrams and bigrams. Each segment was represented by a binary word vector ('1' if the word is in the segment, '0' otherwise). The number of keywords automatically extracted was reduced with feature selection to keep the most important and significant ones (removing less frequent and correlated unigrams / bigrams).

We then trained 3 stages of an automatic annotation tool to identify segments with questions, course *vs.* non-course related questions and the nature of those questions. Overall, 7 classifiers were then trained to annotate the corpus of segments respectively: (1) into question/non-question; (2) into course/non-course related questions (3) for non-course related questions, according to dim 0; (4-7) for course-based questions, according to dim 1 to 4. For each classifier we trained 6 different models : Support Vector Machine (SVM), Generalized Linear Model (GLM), Gradient Boosted Trees (GBT), Decision Tree (DT), K-NN, Naive Bayes (NB) and Rule Induction (RI). All models were evaluated using 10 fold cross-validation on each of step.

Links between questions and success

To answer to RQ2, first we performed four clustering analyses using K-Means algorithm (with k varying between 2 and 10) over four datasets: students who asked questions in GDP5 (N_5= 278 students), GDP6 (N_6= 275), GDP7 (N_7=314) and GDP8 (N_8=287). We performed the clustering using as features for each student the proportion of each question asked in each dimension (e.g. the proportion of questions with value "App" in dimension 1) asked overall. The results reveal that two similar clusters are found in each session of the MOOC, called C1 and C2.

The second step consisted in characterizing the clusters by

Dim0	Categories
Soc	**Soc**ialization
Adm	**Adm**inistrative issues
Exa	**Exa**m/ quiz modality
Tec	**Tec**hnical issues
Res	**Res**sources not found
Too	**Too**ls
Pha	**Pha**tic (has no real value)

Table 3: Coding scheme created (translated from French) from manual annotation to tag non-course related students' questions

analyzing which of the 19 dimensions used to extract them differ significantly. We ran 76 (19 times 4) Mann-Whitney U tests for each dimension for each of the 4 sessions, and adjusted the threshold p-value with Bonferroni correction (adjusted p-value = .0007). Table 4 summarizes the results for the dimensions with a statistically significant difference in at least one of the four sessions. When compared to C2, students in C1 always ask more questions about the exams, less verification questions in particular about concepts from the course. They sometimes also ask more administrative questions (GDP5 and 6), less questions linking two concepts (GDP7) and about how to proceed (GDP8) as well as less questions to deepen their understanding (GDP8).

The third step consisted in characterizing the clusters in terms of attributes not used for the clustering. We ran 4 Chi-square tests for the success and 4 Mann-Whitney U for the final grade. Those tests revealed a statistically significant difference for the final grade for sessions 6 and 8 only ($p = .014$ and $.040$ and $\eta^2 = 0.018$ and 0.010 respectively), with a higher final score for C2, and a higher proportion of students from C2 who obtain their certificate at the end for session 8 only ($\chi^2 = 6.77$, $p = .009$, 79.9% vs. 65.5%).

		dim0$_{exa}$			**dim0**$_{adm}$			**dim1**$_{ver}$			**dim1**$_{dee}$			**dim3**$_{man}$			**dim3**$_{lin}$			**dim4**$_{con}$			**F. Grade**			**Suc.**	**Course**
Cluster	*N*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Q1*	*Md*	*Q3*	*Prop*	*Prop*
$C1_{GDP5}$	189	0	0.25*	1	0	0*	0.33	0	0*	0	0	0	0.18	0	0	0	0	0	0	0	0*	0	47.2	66.3	90.8	0.74	0.20*
$C2_{GDP5}$	96	0	0*	0	0	0*	0	0	0.93*	1	0	0	0.12	0	0	0	0	0	0	0.5	0.67*	1	52.82	82.4	92.22	0.78	0.86*
$C1_{GDP6}$	177	0	0.33*	1	0	0*	0.17	0	0*	0	0	0	0.14	0	0	0	0	0	0	0	0*	0	46.77	56.47*	89.67	0.70	0.22*
$C2_{GDP6}$	98	0	0*	0	0	0*	0	0.50	1*	1	0	0	0	0	0	0	0	0	0	0.50	0.75*	1	51.88	84.61*	93.64	0.78	0.86*
$C1_{GDP7}$	189	0	0.33*	1	0	0	0	0	0*	0	0	0	0.33	0	0	0	0	0*	0	0	0*	0	79.80	91.15	94.70	0.80	0.28*
$C2_{GDP7}$	125	0	0*	0.15	0	0	0	0.50	0.70*	1	0	0	0.17	0	0	0	0	0*	0	0.50	0.67*	1	81.72	91.92	96.20	0.81	0.84*
$C1_{GDP8}$	88	0.67	1*	1	0	0	0	0	0*	0	0	0*	0	0	0*	0	0	0	0	0	0*	0	33.30	84.68*	91.93	0.66*	0.08*
$C2_{GDP8}$	199	0	0*	0	0	0	0	0	0.44*	1	0	0*	0.33	0	0*	0	0	0	0	0	0.33*	1	77.90	88.33*	92.61	0.80*	0.66*

Table 4: Summary of median, first and third quartiles of the variables used for the clustering (* means significant with Bonferroni correction) and for the dependent variables (final grade, proportion of success and proportion of course questions) for each cluster and for each course

We also see that the proportion of course questions is significantly higher for C2 for each session.

Discussion and Conclusion

We have presented a tool to annotate MOOC posts in a more fine-grained manner than usual approaches. When annotating posts across different sessions of a same MOOC, consistent clusters of questions appeared, which are sometimes correlated with the performance. Although course vs. non-course may be enough to help in success prediction, our approach offers a better understanding of the nature of questions from successful vs. unsuccessful students, opening the path to a finer interpretation of what some students are doing wrong. We envision to replicate this analysis on other MOOCs to see if similar patterns can be found.

REFERENCES

1. Omaima Almatrafi, Aditya Johri, and Huzefa Rangwala. 2018. Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. *Computers & Education* 118 (March 2018), 1–9.
2. Yi Cui and Alyssa Friend Wise. 2015. Identifying Content-Related Threads in MOOC Discussion Forums. In *Proc. of the Second (2015) ACM Conf. on Learning @ Scale - L@S '15*. ACM Press, Vancouver, BC, Canada, 299–303.
3. Fatima Harrak, François Bouchet, Vanda Luengo, and Pierre Gillois. 2018. Profiling Students from Their Questions in a Blended Learning Environment. In *Proc. of the 8th International Conf. on Learning Analytics and Knowledge (LAK '18)*. ACM, New York, NY, USA, 102–110.
4. Carolyn Penstein Rosé and Oliver Ferschke. 2016. Technology Support for Discussion Based Learning: From Computer Supported Collaborative Learning to the Future of Massive Open Online Courses. *International J. of Artificial Intelligence in Education* 26, 2 (June 2016), 660–678.
5. Glenda S Stump, Jennifer DeBoer, Jonathan Whittinghill, and Lori Breslow. 2013. Development of a Framework to Classify MOOC Discussion Forum Posts: Methodology and Challenges. (2013), 20.
6. Alyssa Friend Wise and Yi Cui. 2018. Learning communities in the crowd: Characteristics of content related interactions and social relationships in MOOC discussion forums. *Computers & Education* 122 (July 2018), 221–242.