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

Fatima Harrak[1], François Bouchet[1], Vanda Luengo[1] and Rémi Bachelet[2]
[1]Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, Paris, France
[2]Centrale Lille, University of Lille, France

INTRODUCTION

In MOOCs, **discussion forums** are a key feature and although several works have tried to show the impact of **categorizing students' posts** [2], in terms of content-relatedness [3], or urgency [1], they rarely look into the posts' detailed content.

Hypothesis : analyzing more finely the content of MOOC posts would help in predicting students' success. We investigate the following questions:

1. Can we reliably **annotate** questions extracted from MOOC forum posts according to a fine-grained multi-level **coding scheme**?
2. Is there a consistent **relationship** between **students' questions** and their **performance** in the MOOC?

DATASET

Context: French MOOC on project management called **GDP** (French acronym for project management)

Data type:

- Log data: raw Canvas logs
- Forum data: threads created by the pedagogical team answering technical or administrative issues, about homework or **course content**, among others.

Data size: 4 sessions (5 to 8) from 2015-2016

	#students	#posts	#unique posters
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

For each session we extracted:

- the students' posts in **course related** topics,
- the **final grade** (out of 100)
- students' **success** (grade above 50 out of 100).

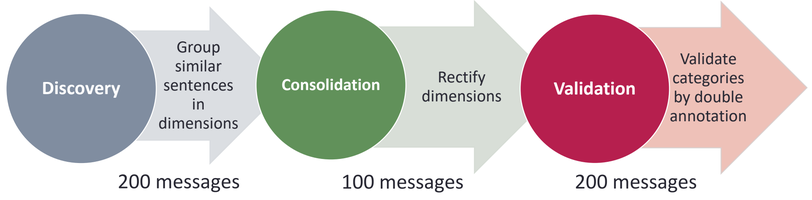
RESEARCH METHODOLOGY

To answer to the research questions, we have followed 3 successive steps:

1. **Manual process** to extend an existing coding scheme
2. **Automatic annotation** of students' posts using several classifiers in cascade
3. **Clustering** over four sessions and then characterization of the obtained clusters

CODING SCHEME DESIGN

We took a sample of **500 messages** from threads which are course related of 4 sessions to apply 3 different categorization steps (as proposed by Harrak et al. in other context of study [4]).



PROPOSED CODING SCHEME

Dim1	Question type
Ree	Re-explain / redefine
Dee	Deepen a concept
Ver	Validation / verification
Dim2	**Explanation modality / question subject**
Exa	Example
Sch	Schema
Cor	Correction
Dim3	**Explanation type**
Def	Define
Man	Manner (how?)
Rea	Reason (why?)
Rol	Roles (utility?)
Lin	Link between concepts
Dim4	**Verification type (optional)**
Mis	Mistake / contradiction
Kno	Knowledge in course
Exp	Expected knowledge

Table 2. Coding scheme of course-based students' questions

Example of course-related question:

"Could you detail the differences between layers and underlays?"

[Dee, 0, Lie, 0]

Dim0	Categories
Soc	Socialisation
Adm	Administrative issues
Exa	Exam / quiz
Tec	Technical issues
Res	Resources not found
Too	Tools
Pha	Phatic

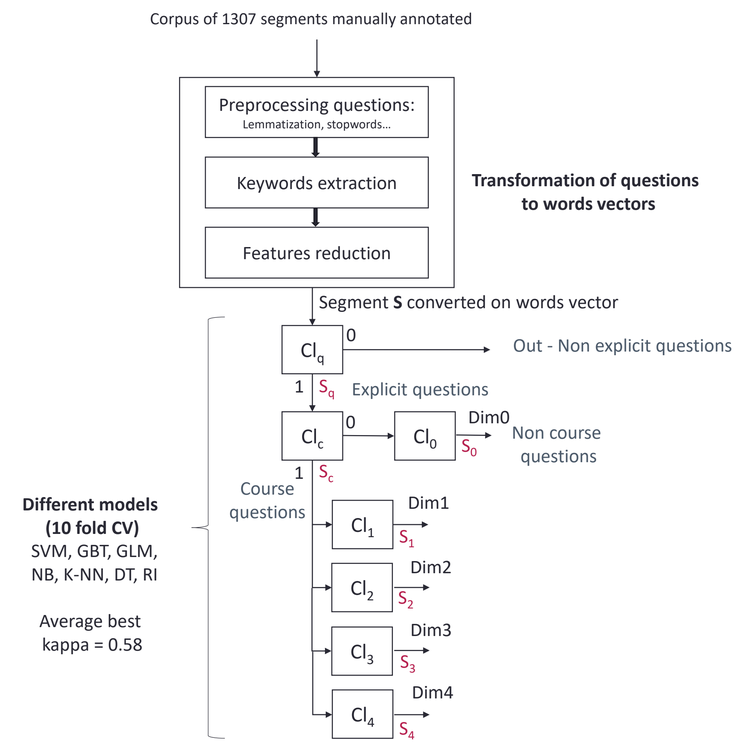
Table 3. Coding scheme of non-course related students' questions

Example of non-course related question:

"Where to find the PDF version of the course?"

[Res]

AUTOMATIC ANNOTATION



STUDENTS' CLUSTERING

We performed four clustering analyses using K-Means algorithm (with k in [2..10]) over four datasets: students who asked questions in GDP5 (*N5* = 278 students), GDP6 (*N6* = 275), GDP7 (*N7* =314) and GDP8 (*N8* = 287). We ran Mann-Whitney U and Chi-square tests to reveal a statistically significant difference for each cluster (adjusting for multiple testing using Bonferroni's correction).

We used as features for each student the proportion of each question asked in each dimension overall. We obtained 2 similar clusters in each session of the MOOC:

	Cluster 1	Cluster 2
Final Grade	lower (GDP6&8)	Higher (GDP6&8)
% Successful	Low (GDP8)	High (GDP8)
Exa	More (all)	Less (all)
Adm	More (GDP5&6)	Less (GDP5&6)
Ver	Less (all)	More (all)
Dee	Less (GDP8)	More (GDP8)
Man	Less (GDP8)	More (GDP8)
Lin	Less (GDP7)	More (GDP7)
Con	Less (all)	More (all)
% Course questions	Lower (all)	Higher (all)

Table 4. Summary of variables with statistically significant differences between the 2 clusters across the sessions

CONCLUSION

- This work allows us to annotate MOOC posts in a more fine-grained manner than usual approaches
- We have found consistent clusters of questions which are in some cases correlated with the performance
- Our approach offers a better understanding of the nature of the questions of successful vs. unsuccessful students, opening the path to a finer interpretation of what some students are doing wrong

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] 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.

[3] 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*, 221-242.

[4] Fatima Harrak, François Bouchet, Vanda Luengo, and Pierre Gillois. 2018. Profiling Students from Their Questions in a Blended Learning Environment. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK '18). ACM, New York, NY, USA, 102–110.