Semi-automatic annotation of MOOC forum posts

Weizhe Liu, Łukasz Kidziński and Pierre Dillenbourg

CHILI Laboratory, École polytechnique fédérale de Lausanne,
RLC D1 740, CH-1015 Lausanne
http://chili.epfl.ch/

Abstract. Massive online open courses' (MOOCs') students who use discussion forums have higher chances of finishing the course. However, little research has been conducted for understanding the underlying factors. One of the reasons which hinders the analysis is the amount of manual work required for annotating posts. In this paper we use machine learning techniques to extrapolate small set of annotations to the whole forum. These annotations not only allow MOOC producers to summarize the state of the forum, but they also allow researchers to deeper understand the role of the forum in the learning process.

1 Introduction

One of the main differences between MOOCs and university courses is the limited social interactions. Researchers tried to investigate this component, suggesting for example MOOC study groups [2]. Nevertheless, for the moment, a forum remains the central channel of collaboration.

Forum activities positively correlate with grades and retention [3]. However, little research has been done to analyze these relations on a more granular level than active or passive activities. Deeper analysis requires tedious manual annotations, preferably done by several judges independently. From the research perspective, this significantly increases the cost of the analysis, whereas from practical perspective, in case of courses with thousands of users, it is not feasible to track all the posts annotations dynamically during the course. We argue that recent machine learning improvements allow us to address this problem and benefit in both cases.

We distinguish two machine learning approaches to the problem: unsupervised and supervised learning. The unsupervised learning allows to group posts together according to imposed measures. Semantics of the groups are not determined by the algorithm and often they are difficult to interpret. Conversely, the supervised learning is based on ground truth - a set of post annotated manually in which semantics are predefined. It allows us to extend known annotations to a larger set. Unsupervised techniques were recently successfully applied for clustering dialogues in MOOC forums [4]. In this study we employ the supervised approach and refer to the technique as a *semi-automatic annotation*. We present the dataflow in Figure 1.

Semi-automatic text classification was successfully used to analyze political forums [6] and technical forums [9] [8]. In addition, both written and spoken conversations were analyzed [7]. Although the latter is outside the context of our work, still the techniques prove to be universal.

The semi-automatic annotation can support the work of MOOC practitioners and other researchers. As an example application of our technique, we check to what extent we can identify threads important for the learning process. This could allow us, for example, to support teaching staff by alarming them about important questions, which have not yet been addressed. Another possible application is classification of users according to their behavior and investigating if certain activity patterns reflect their performance in the course.

Our techniques can also support practitioners in automatic assessment of measures already established in literature, which are costly to implement manually in practice. This includes scoring posts in discussion forums, based on the overall contribution of the post to the thread or subject [11] [12], supporting forum management [1], measuring emotional content of posts [] [10] or engagement of students [5]. Finally, annotations provide another dimension for the analysis of designs choices, and may extend the previous results of Coezee et al. [3].

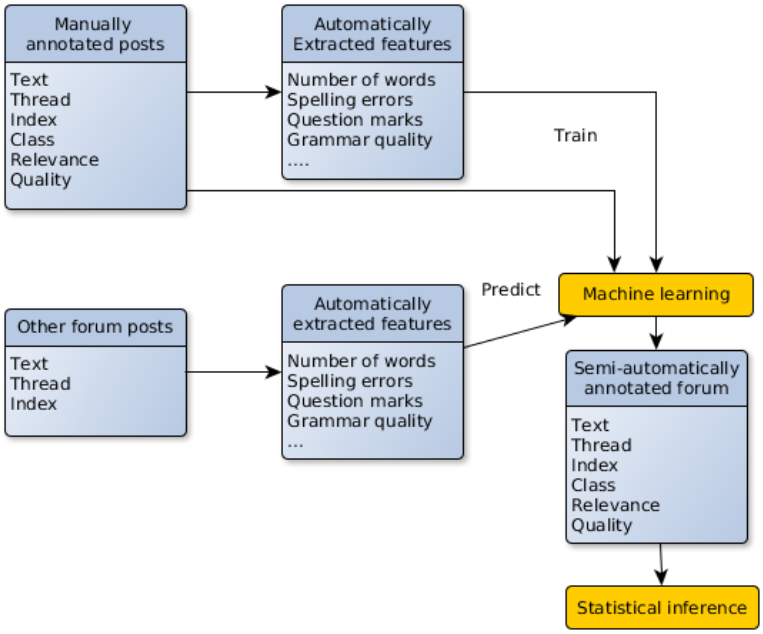

Fig. 1. Instead of manually annotating the whole forum we annotate a small part and employ machine learning techniques. This allows us to draw conclusions from the whole annotated forum.

Contributions of this work are two-fold. First, we suggest methodology for automatic annotation of forum posts, which can be beneficial for many MOOC stakeholders. Second, we use estimated annotations for large scale analysis of performance of students participating in MOOC forums and we answer research questions: Do students who ask questions perform better than students who answer? To what extent the quality of the post reflects students' performance in the course? How can we support work of teaching assistants using automatic annotations?

2 Dataset

We analyze forum data from the Scala MOOC given by the EPFL in 2014. Our dataset consists of 4316 posts of 1336 different users. Apart from the content of the post, we also have the id of the thread, id of the author, timestamp of posting, position in the thread and the title of the thread. In order to reduce the impact of the language factor, we only consider posts written in English. Moreover, we know the final grade of each user, or we have information that they dropped out.

3 Post annotation

As we want to classify and compare forum behaviors, like *student asking questions*, *student answering questions*, etc., we suggest several **classes of posts**. We use classification scheme introduced by Sridhar et al. [6], altered for MOOC context. For possible values of this feature we refer to Table 1.

Since one objective of automatic annotation in our study is the detection of important posts which were not yet properly addressed, we choose to assess the *importance* of the content in the relation with the course. To that end, we introduce two indicators: *Relevance* and *Comprehensibility*. These notions are subjective to this study and can be altered in other specific applications.

Relevance describes how closely related the content of the post is with the content of the course. We distinguish 5 levels of relevance:

1. *off-topic* - student has no intention to be even close to the content of the course,
2. *slightly irrelevant* - the relation is very superficial and main point of the post is off-topic,
3. *neutral* - there is some relation to the topic, but vague and not clear,
4. *relevant* - there is a clear intention to understand/solve some problem,
5. *highly relevant* - the student demonstrates understanding and his own tries to solve the problem

The relevance varies only for three post classes: *Question*, *Answer* and *Clarification*. In other cases the relevce is implicit, for example, *off-task* posts are *off-topic* and there is no room for variability prediction.

Comprehensibility expresses how easily readable is the post. We consider three levels of this feature

Class	Description
Question	Requests of information about the course
Answer	Attempt to answer the question posed in the thread
Clarification request	Request for more details/clarifications concerning the solution
Clarification	Follow-up answer to the request or additional clarification to the answer even without the request
Positive feedback	Student's (not necessarily question author) positive feedback (gratitude etc.)
Negative feedback	Student's negative feedback (solution does not work, have errors etc.)
Off-task	Any form of spam or misplaced text, not relevant to the thread

Table 1. Post classes ordered according to their *importance* in the context of our study. If a post belongs to several classes at once, we classify it to a *more important* class.

1. *misleading* - unclear or misleading language, where it is not clear what is the message,
2. *neutral* - clear and accurate statement,
3. *high comprehensibility* - student states all the necessarily details of the problems and/or his own attempts to the solution.

4 Automatic annotation

Our goal is to predict variables introduced in Section 3, given a small set of manual annotations. To this end, we automatically extract features presented in Table 2 from the content of the forum and use machine learning techniques which were trained on the manually annotated dataset. A set of *relevant words* was determined from the content of lecture slides and posts, manually by listing words which appear the most often and extracting those which are specific to the course. These words were used for `RelWords*` features. `GrammarQuality`, `ErrSpell` and `ErrGram` were automatically extracted using R text mining `tm` package.

The basic set of features from Table 2 can be still extended in order to achieve more accurate classification, however, in the context of our research questions the results were already satisfying.

The extracted features were used for training machine learning algorithms using R package `CARET`. As the training set we use 100 randomly selected posts, annotated manually by one student. This annotation introduces additional error, however, for the purpose of this study we treat these annotations as the ground truth.

Name	Description
`NumWords`	Number of words
`RelWords`	Number of relevant words
`RelWordsRatio`	Ratio of relevant words
`NumSent`	Number of sentences
`ErrSpell`	Number of spelling errors
`ErrGram`	Number of grammar errors
`NumQMarks`	Number of question marks
`Index`	Index in the thread (1 - for the first post in the thread)
`NumWordsTitle`	Words in the thread title
`RelWordsTitle`	Relevant words in the thread title
`RelWordsTitleRatio`	Ratio of relevant words in the title
`GrammarQuality`	Correctness of grammar (1-10)

Table 2. Features extracted from the posts.

4.1 Measures of accuracy

In the classification of posts we measure accuracy of prediction by out–sample Cohen's κ. For relevance and comprehensibility, since both indicators can be treated as ordinal, we will use regression models to predict the values. We measure accuracy of these models by root mean squared error (RMSE). For example, for relevance prediction, we define

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} \|R_j - \hat{R}_j\|^2},$$

where R_j is the relevance of the j-th post, $\hat{R}_j$ is the predicted relevance and n is the number of posts. RMSE approximates average deviation of prediction from the observed value.

Both measures were estimated by cross-validation. In each case, we repeated the following procedure 5 times: We take randomly 80% of the training observations, we predict remaining 20% and we compute the measure. The average of 5 measurements serves as the estimator of the measure.

4.2 Results of the semi-automatic annotation

In classification task we implement several machine learning methods including *Multiclass Logistic Regression*, *Bayesian Model*, *Random Forest Model*, *Support Vector Machine* along with *Kernel Method* to maximize Cohen's κ, and we find that *Random Forest by Randomization* performs the best in our context, with $\kappa = 0.57$.

We use the 10 extracted features to predict the *Relevance* and take root means squared error (RMSE) as a measure of accuracy. After testing several regression methods, we employ *SVM with Linear Kernel* which yields the lowest

RMSE equal 0.96. Finally, in the same way we predict *Comprehensibility*, in this case we chose *Penalized Linear Regression* as the most efficient method, giving RMSE is around 0.54. Other values are presented in Table 3 and Table 4.

Method	Cohen κ
SVM	0.42
Bayesian Model	0.40
Logistic regression	0.35
Random Forests	0.57

Table 3. Cohen's κ of various methods used for classification. The larger the better.

Method	Relevance	Comprehensibility
Linear regression	0.98	0.58
Penalized linear regression	0.98	0.54
Neural network	1.03	0.69
SVM	0.96	0.56

Table 4. RMSE of regression of Relevance and comprehensibility using various methods.

5 Applications of annotations

In this section we discuss possible applications of semi-automated annotator, both for practitioners and for researchers. First, the direct application is the prediction of *important* posts, where, for for, importance is defined by relevance to the subject and high comprehensibility. Teaching assistants, instead of reading posts one by one, could generate a list of important posts which have not yet been answered. Second, we can analyze performance of students with certain posting patterns. In this section we report the latter analysis.

We investigate if students who ask more questions achieve better results than those who answer more questions. To this end, we compare two subgroups of students, those whose questions account for more than 80% of their posts, those whose answers account for more than 80% of their posts. We find that students who answer more questions perform better than those who ask more questions ($T = -4.9829, df = 541, p < 0.01$), as depicted in Figure 2.

We also analyze the *Relevance* of posts as a predictor of students' performance. Hence, we also set two groups of students, the first group consists of students with average relevance above 4 and the second group consists of students with average relevance below 2. We compare distributions of the results within these two groups of students and present it in Figure 3. As the Figure

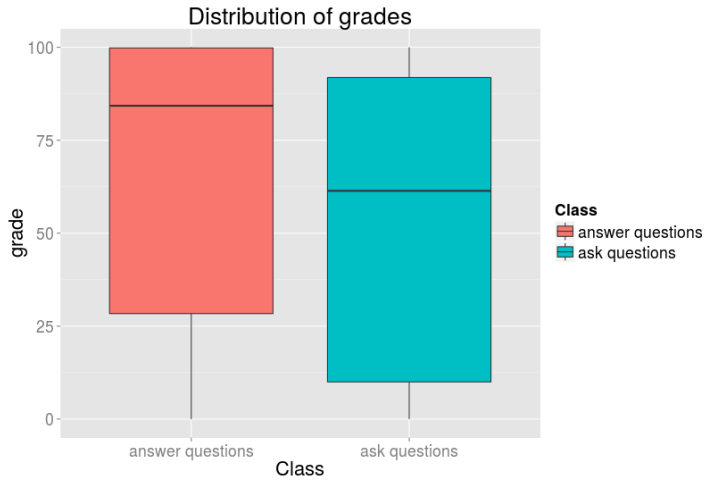

Fig. 2. Mean grades between students who ask more questions and answer more questions.

shows, the students with low relevance obtain much better performance than those with high relevance ($T = -5.6875, df = 146.35, p < 0.01$).

Note that we considered the extreme cases, i.e. relevance smaller than 2 or larger than 4. In case of more relaxed thresholds 2.5 and 3.5 the effect size is smaller but still significant at $\alpha = 0.05$.

As the result is very unintuitive, we analyze manually the posts which were predicted to be less relevant. We find that indeed these posts are not close to the content of this course. Instead, they often concern social interactions like finding friends. Questions concern, for example, nationality of other students (*"where are you from?"*). The posts which were predicted to be close to the content of the course are highly related to the lessons. This indicates that a variable explaining *social interactions* could be a good indicator of performance. In future studies, we suggest emphasizing this social aspect instead of *low relevance*, leading to more intuitive interpretation.

For finding the relation between *Comprehensibility* and the final grades, we divide students into different levels of average post comprehensibility. We define low comprehensibility as lower than 1.5 and high comprehensibility as larger than 2.5. Although the results of students with high comprehensibility posts may appear to have smaller variance in Figure 4, there is no statistical evidence for such claim ($F = 1.222$, df $= 163$, $p = 0.3575$). Moreover, the mean grades in these two groups are not significantly different.

Deeper investigation of Comprehensibility indicator reveals that our notion of *Comprehensibility* is not relevant in the context of programming courses, particularly because posts containing chunks of source code are classified as low comprehensibility, whereas in the context of programming this code substan-

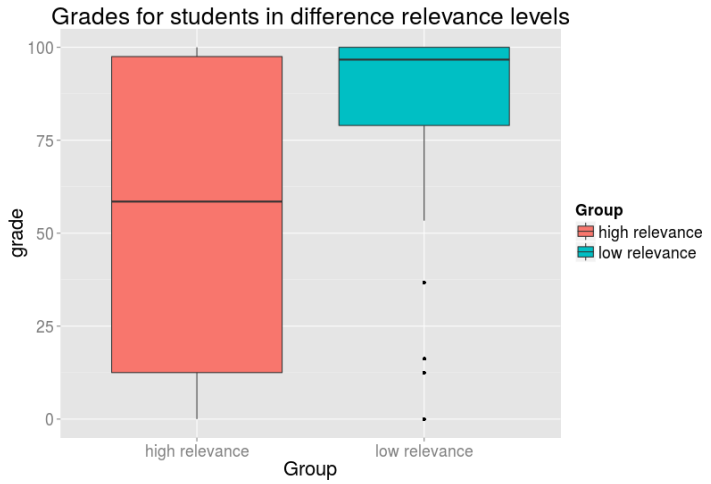

Fig. 3. Box-plot of grades of students with different average relevance level.

tially improves readability. This problem can be solved by introducing features based on the detected source code, as well as ground truth according to which posts with code are classified higher.

Other low comprehensibility posts are those with very short sentences, just one or two words in each post. Capturing the variability of these short posts may require techniques beyond the introduced framework. Note that in our analysis we rely on the assumption that posts treated as independent entities contain information sufficient for classification and regression. However, a post containing only "No" is very strongly context-dependent. Depending on the task we can either remove these posts from the analysis or seek for more complex models, which take into account correlations within a thread.

6 Limitations

Through the data analysis process, we found that the seven classes in our model do not cover all possible forum interactions. For example, some students just share their thoughts about the course and they do not ask any direct questions. This can be addressed by the introduction of additional classes.

The 10 features we extract directly from the posts are not enough for distinguishing some of the classes. For example, in our setup *Clarification Request* and *Negative Feedback* have similar values of predictive features. Depending on the application we could either join the classes to a general one or introduce more precise features, e.g. an indicator of phrases "doesn't work", "sorry", etc.

Some students tend to cite the post before answering it. Posts of this kind may be misclassified as questions while they are actually answers. Again, depending

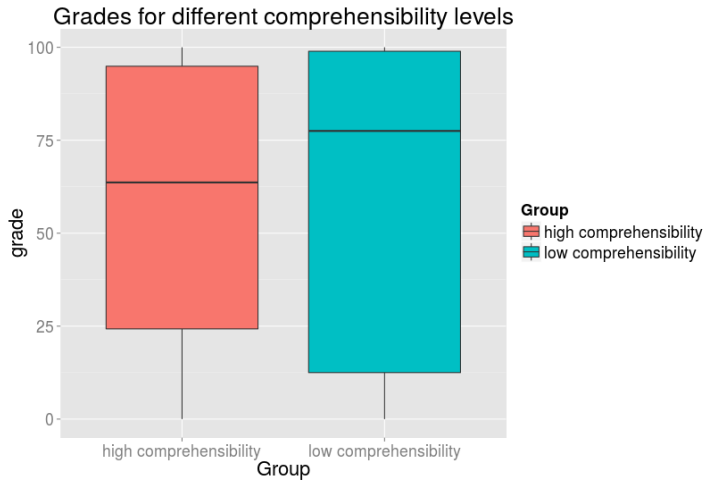

Fig. 4. Box-plot of grades of students with different average comprehensibility level.

on the application and the forum structure, another predictive feature could be introduced.

Finally, posts were annotated just by one judge, and these annotations were taken as ground truth, which can lead to significant inaccuracy. Larger dataset of annotated post and annotation from independent judges would allow us to understand deeper technical aspects of the problem.

7 Conclusion

In this project, we used machine learning techniques to annotate posts and answer questions related to the users' behavior in forum. We introduced a methodology which allows to answer research questions given only a limited number of annotations. We find that students who answer more questions achieve higher grades than those who ask more questions. Moreover, students who are more social in the MOOC performed better. We also find that the *Comprehensibility* measure, based merely on the elementary features and small set of annotations, should be altered for programming courses. This measure in conjunction with relevance can simplify the work of teaching assistants in a MOOC. Although our elementary methodology already yields satisfactory results, many improvements can be incorporated for specific applications of semi-automatic annotations.

References

1. Bhatia, S., Biyani, P., Mitra, P.: Classifying user messages for managing web forum data (2012)

2. Blom, J., Verma, H., Li, N., Skevi, A., Dillenbourg, P.: Moocs are more social than you believe. eLearning Papers (2013)
3. Coetzee, D., Fox, A., Hearst, M.A., Hartmann, B.: Should your mooc forum use a reputation system? In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. pp. 1176–1187. CSCW '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2531602.2531657
4. Ezen-Can, A., Boyer, K.E.: Unsupervised classification of student dialogue acts with query-likelihood clustering. In: Educational Data Mining 2013 (2013)
5. Ramesh, A., Goldwasser, D., Huang, B., Daume III, H., Getoor, L.: Modeling learner engagement in moocs using probabilistic soft logic. In: NIPS Workshop on Data Driven Education (2013)
6. Sridhar, D., Getoor, L., Walker, M.: Collective stance classification of posts in online debate forums. In: ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media (2014)
7. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics 26(3), 339–373 (2000)
8. Wang, L.: Thread and post classification over technical user forum data
9. Wang, L., Lui, M., Kim, S.N., Nivre, J., Baldwin, T.: Predicting thread discourse structure over technical web forums. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 13–25. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145435
10. Wen, M., Yang, D., Rosé, C.P.: Sentiment analysis in mooc discussion forums: What does it tell us? Proceedings of Educational Data Mining (2014)
11. Xue, X., Jeon, J., Croft, W.B.: Retrieval models for question and answer archives. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 475–482. SIGIR '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1390334.1390416
12. Zhou, L., Hovy, E.H.: On the summarization of dynamically introduced information: Online discussions and blogs. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. p. 237 (2006)