Investigating Students’ Learning in Online Learning Environment

Lavendini Sivaneasharajah, Katrina Falkner, Thushari Atapattu
The University of Adelaide
{lavendini.sivaneasharajah, katrina.falkner, thushari.atapattu}@adelaide.edu.au

ABSTRACT

Due to the increasing interest in the online learning environment, particularly in Massive Open Online Courses (MOOCs), predictions and education data mining have rapidly gained prominence in education studies over the past decade. The massive amount of student data available in MOOC platforms enables us to gain insight into students’ learning behaviours. Therefore, this paper outlines the doctoral work that explores the idea of ‘student roles’ and their linguistic changes to analyse the students’ learning behaviours in MOOCs. A multi-class classifier has been built to identify user roles (e.g. information seeker, information giver) with 82.30% F-measure. Preliminary results on linguistic experiments demonstrate, distinguish linguistic behaviours can be observed in different user roles. The outcome of this research study will contribute to a learning model that can be used to understand students’ learning process.

Keywords

MOOCs, Discussion forums, User Role, Natural Language Processing, Machine Learning.

1. INTRODUCTION

Learning Analytics has been gaining high popularity in recent years among research scholars due to the challenges it imposes; two of which are increasingly complex, large-volume data and heterogeneous data. Integrating several sources of data that are generated during learning activities is of major need for the education sector to provide timely enhanced services to both students and instructors. Integrating and analysing student data can contribute immensely towards reducing dropout rates, timely instructor interventions, and many other [4].

In the 21st century, students are more exposed to Massive Open Online Courses (MOOCs) and online learning environments as they believe it is more beneficial than the traditional learning environment such as flexible study hours, availability for everyone [7]. As many of the MOOCs are freely available for students, it draws the interest of thousands of learners. However, accessing the success rate of a student learning in online platforms has become difficult as students enrol for varying purposes. Knowing that students may enrol in courses for other purposes, we need to explore other perspectives of learning success beyond completion.

MOOC contains many types of resources to support students in their learning activities. These elements can be categorised as videos, lecture series, reading materials, quizzes, assignments, discussion forums etc. According to Anderson [1], discourse enables the learner to come up with their own reasoning and logical thinking by communicating with others. Thus, investigating discussion forums will help researchers to understand the actual situation of the students in the learning lifecycle.

The overarching aim of this doctoral work is to understand students’ learning with time within MOOCs. To this end, the research mainly focuses on examining user role transformation and linguistic change that occurs in discussion forums with time. We believe analysing these roles and associated linguistic changes will eventually result in a deeper understanding of the student’s learning lifecycle. Further, this research will also investigate the influencing factors (e.g. course structure, learners’ demographic) that influence these observable features (i.e. student role, linguistic expression) and their correlations.

To achieve the aforementioned aim, our investigations are driven by two main research questions (RQ):

RQ 1: Can student role and linguistic expressions be used to understand student learning? (1. How to build a predictive model that predicts students’ roles in an online learning environment? 2. How to track the linguistic change of each student roles in the online learning environment?).

RQ 2: To what extent students’ learning is affected by external factors? (1. What are the external factors that affect these transformations (user role and linguistic change)? 2. What are the correlations between external factors and these transformations?).

The contribution of this doctoral work includes a predictive model that leverages linguistic-only features to predict student roles in discussion forums. Further, this doctoral work develops a linguistic framework to understand students’ learning. This demonstrates distinguish linguistic behaviours of different student clusters in discussion forums. Moreover, identifying the external factors such as course structure, learners’ demographic that affect students’ learning and their correlation.

2. RELATED WORK

2.1 Post classification and role identification in discussion forums

User role classification is grounded by post-classification methodologies that prevail in the existing literature. In other words, post-classification is the foremost step that needs to be carried out in order to identify the user roles in discussion forums. With the examinations on speech acts by Searle [9], there are

Lavendini Sivaneasharajah, Katrina Falkner and Thushari Atapattu "Investigating Students' Learning in Online Learning Environment" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 806 - 809

several post-classification methodologies have been introduced to the research community.

While prior studies classify forum posts into different categories such as question, answer, solutions, Hecking et al. [6] have carried out post-classification by generalising the categories that prevail in the existing studies [2]. The study presents three different classes namely: information seeking, information giving and other. Hecking et al. [6] achieved 70% accuracy using content-related features (e.g. phrases – "need help or helps you") and contextual features (e.g. position in the thread, number of votes) for classification purposes. However, relying on contextual features for forecasting is not feasible in a real-time system as these contextual features changes with time. And predictions can only be made at the end of the course as they occur during the course. Therefore, our study aims to build a predictive model for discussion forum classification using linguistic-only features while eliminating contextual and structural features.

2.2 Linguistic change in online communities

For decades, researchers in the linguistic discipline have explored the language change in many different spheres starting from historical linguistics to sociolinguistic. Research scholars believe exploring temporal changes in user's language will provide useful insights to research communities. Linguistic research has taken various paths with time to exhibit correlations between the linguistic and other aspects such as historical change, community norms, user lifespan etc.

The work by Nguyen et al. [8] identifies the relationship between community membership and language use. According to their findings, forum specific jargons and informal linguistic style can be observed in long-term participants' discourse. Dowell et al. [5] have conducted a study on MOOC data to identify the conversion in learner's language and discourse characteristic with time. However, the research did not investigate the linguistic changes associated with each user role. It is said that learner's language changes with time, especially discourse in discussions forums will be topic-oriented and reflective of deep learning with the consequent offerings of a course [5]. Nevertheless, investigating linguistic change for a student role has not been addressed. Even though preliminary work on linguistic change has been conducted in other online communities, there is a lack of work conducted in MOOCs.

3. METHODOLOGY

This doctoral work will be conducted in two main phases namely:

1. Pilot study - Identifying potential features from discussion forums.

2. Building machine learning model - Implementing a model to understand student learning. The phase two will be further divided into three sub tasks to address aforementioned research questions as follows:

Task 1: Developing predictive model to identify user roles (IG/IS and O) in discussion forums.

Task 2: Developing a machine learning model to track linguistic change.

Task 3: Identifying external factors and their correlations with user roles and linguistic expressions.

4. EXPERIMENTS AND RESULTS

The study collected 9,497 user posts from 923 users from the AdelaideX[1] 'Introduction to Project Management' and 'Risk Management for Projects' courses offered in 2016 and 2017 respectively. The current study was conducted using 6000 posts from 'Introduction to Project Management'. Two independent human evaluators carried out a manual annotation with a high inter-rater agreement (Cohen's kappa = 0.925) and annotated the user posts as information seeker (IS), information giver (IG) and other (O).

4.1 How to build a predictive model that predicts students' roles in an online learning environment? (RQ1)

A multi-class classifier was built to predict user roles (IG, IS and O) for a given forum post using discourse features and linguistic features. The features were extracted using Pennebaker's Linguistic Inquiry and Word Count (LIWC) tool[2] which generates different linguistic measures for an input text. The study selected sixteen optimal features using Recursive Feature Elimination with Cross-Validation feature selection technique.

We implemented following multiclass classifiers with different sets of algorithms using Weka: Naïve Bayes, Random Forest, Simple Logistic Regression, Logistic Regression and Sequential Minimal Optimisation (SMO). All these classifiers were tested using 10 Fold Cross-Validation to assess the accuracy. Among these, the Random Forest classification model performed best with 82.30 of F measure.

Further, we also fine-tuned the parameters for Random Forest classifier using the scikit-learn library (RandomizedSearchCV and GridSearchCV). The results show that Random Forest classifier performs at its best in the following parameter setting: n_estimators':400, 'min_samples_split':10, 'min_samples_leaf': 4 and max_depth': 70.

Table 1: Results of classifier performance

Classifiers	Accuracy	Precision	Recall	F1	Cohen's Kappa
Naïve Bayes	71.28	74.40	71.30	71.00	0.5117
Random Forest	82.17	82.30	82.20	82.20	0.6955
Simple Logistic	79.35	79.60	79.40	79.40	0.6473
Logistic	79.43	79.70	79.40	79.50	0.6498
SMO	74.80	76.50	74.80	75.30	0.5770

The work by Hecking et al. [6] is the only existing work in our workspace that classify the discussion forums posts as information giving, information seeking and other. They have achieved an overall of 71.5 F-measure for IS and IG class while obtained an average of 70 and 66 for precision and recall respectively across all three classes. With 82.30% of F-measure, we have demonstrated that analysing the language of post content itself is sufficient to predict user roles. Therefore, it is evident that linguistic features have a high impact on user role prediction in discussion forums.

[1] https://www.edx.org/school/adelaidex

[2] https://liwc.wpengine.com/

4.2 How to track the linguistic change of each student roles in the online learning environment?

An important component of this research study is to propose a linguistic framework that can exhibit the linguistic characteristic of different student clusters that will be identified by this study. Afterwards tracking their changes associated with each student role. To achieve this, we carried out linguistic experiments to identify distinguishing characteristics between these student roles.

We started with a simple word count and performed the one-way analysis of variance (ANOVA) for different student roles. The analysis shows that the mean value of information giver is higher than information seeker, and there is a significant difference between the mean values ($p<0.005$). The results indicate that information giver tend to use more words when reflecting their thoughts than the information seeker in discussion forums.

Using a similar approach, we computed the frequency of n-grams in given user posts. We created a vocabulary list using n-grams from lecture transcripts. Then, the study computed the lexical frequency profile (LFP) for each user role. We created a Phrase Matcher Object and applied the matcher object on each user post to extract the keywords. For a given user, the number of keywords used in information giving post increases – reach an optimal number and decreases with time, whereas for information seeking post it increases/decreases with time — also, a minimum level of changes observed in other user post. Moreover, information giver uses more keywords from the lecture transcript than information seeker and other. Further, there is a considerable amount of drop-in 'other' user role. Figure 1 shows the Lexical Frequency Profile for a sample of five users with time.

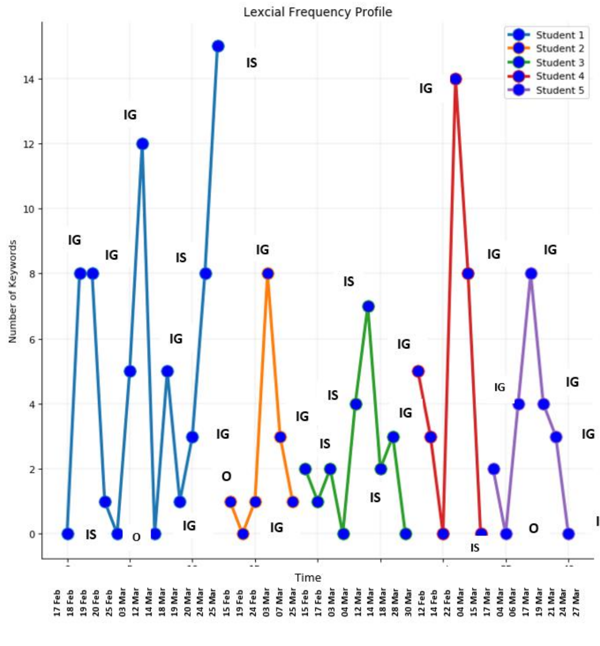

Figure 1: Lexical frequency profile across user role

Information embeddedness is one of the key elements that contribute toward student learning. This study attempts to find the level of information embeddedness using clause extraction. Clause extraction has been used to determine the relationship between the clauses per sentence and language development. We develop a novel approach in which clauses have been extracted from the parse tree using a rule-based approach. A pipeline is being built with Part-Of-Speech (POS) tagging using Stanford CoreNLP[3] to get the basic interpretation of a student post. Tree Annotation is used to extract a parse tree for a given sentence. Initially, clause-level tags (e.g. SBAR) and word-level coordinating conjunction (e.g. CC) have been extracted from the parse tree. Then, we implemented a rule-based approach to extract the number of clauses.

According to Crossley et al. [3], discourse complexity can be measured by any given reading level measures. Therefore, we used Flesch-Kincaid reading level measure to explore discourse complexity with time for each user. Figure 2 demonstrates the discourse complexity for five students with time. The results indicate that if a particular user role can be seen in consecutive posts the level of complexity increases/decreases with minimum change and when there is a role change (e.g. IS → IG or IG → IS or O → IG) there is a dramatic change in discourse complexity.

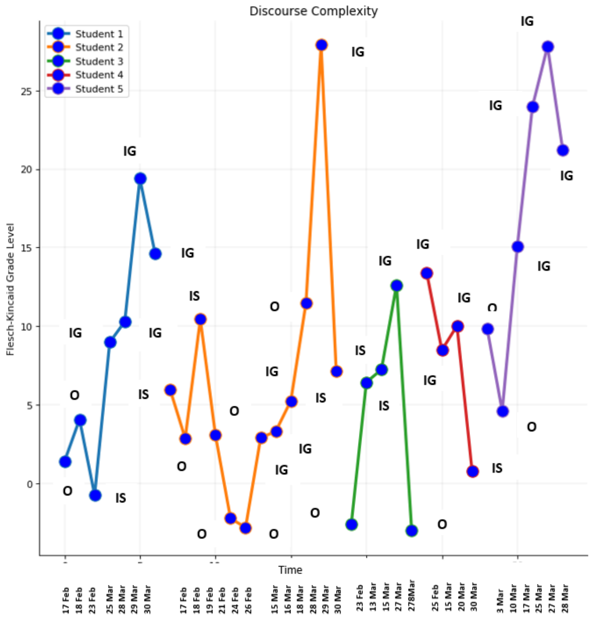

Figure 2: Discourse complexity across user roles with time

An initial exploratory analysis was performed on topic modelling using state of the art topic modelling technique known as Latent Dirichlet Allocation (LDA). We try to identify the topics that have been discussed in each user post and lecture transcripts. Figure 3 shows the percentage of each topic discussed by information givers and information seekers. According to the analysis, information givers are more involved in discussing the latter part of the course topics than information seekers while information seekers show interest towards the beginning of the lecture content. In future, further analysis will be performed to discover the reasons behind this observed trend.

Moreover, we calculated the affective state of each user posts using LIWC tool. Affect features measures the positive and negative sentiment and more specific emotion such as anger, anxiety and sadness. The results of one-way analysis of variance (ANOVA) show that information seekers express more lexical

[3] https://stanfordnlp.github.io/CoreNLP/

semantics associated with affective state than information givers. Likewise, we would like to perform several other linguistic experiments to develop a linguistic framework that will demonstrate the linguistic characteristics of different student clusters in discussion forums.

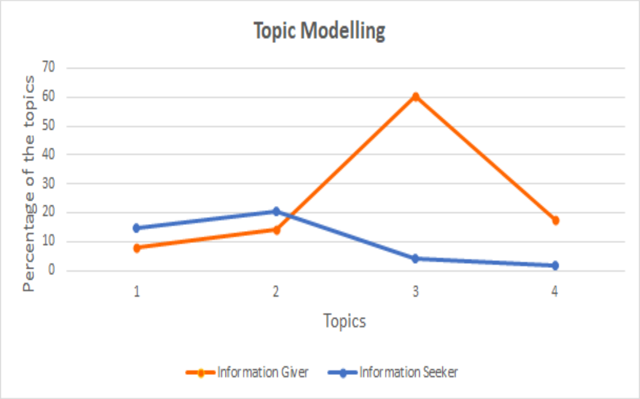

Figure 3: Topic modelling across user role

With the insight gained from existing experiments on user role and linguistic changes, our next step will be to predict the student grading using linguistic-only features. We have done a preliminary analysis on how to divide the student grading into different categories (e.g. pass, fail) and predicted student grades. However, further analysis needs to be performed on feature selection and deploying several other machine learning algorithms to fine-tune the obtained results.

5. CONCLUSION AND FUTURE WORK

The aim of our doctoral work is to understand student learning in MOOCs by investigating user roles and their associated linguistic change. As an initial stage, we have presented a multi-class user role classification in MOOC discussion forums using linguistic-only features with the intention of eliminating the drawbacks (e.g. contextual features) that exist in previous studies. Our model performed well compared to the baseline model, with 82.30 % of F-measure.

As future work, we try to integrate this classification with content and non-content user posts. Thus, it results in a novel classification on user role classification in MOOC discussion forum. On the other hand, our linguistic study gives us a clear differentiation of linguistics aspects associated with each role. Further, we hope to do a meticulous analysis to explore these patterns in future with the intention of discovering the possible reasoning behind the observed trends. Further analysis will be conducted to identify the discourse measures that can contribute to understanding student learning. The study would also like to explore diverse methods/techniques that can discover correlations between these linguistic measures and students' learning in MOOCs. Understanding how these linguistic measures can contribute directly/indirectly to students' learning will help us to propose novel methods to understand students' learning in an online learning environment. In addition, experiments will be performed to identify the correlations between the external factors (e.g. course structure, assignment deadlines) and user role transformations.

As a proof of concept, our technique demonstrated the potential of identifying the linguistic behaviours for each user role. This novel approach holds a great promise for user role classification and the associated linguistic behaviour in MOOC discussion forums. Additionally, we believe that tracking these role changes and associated linguistic changes will help to understand the student learning in MOOC discussion forums. Thus, this doctoral work, will eventually try to find an answer to 'are students' really learning from MOOCs?'

6. ADVICE SOUGHT

For this doctoral consortium, the study would like advice regarding the following concerns mainly focusing on linguistic study:

1. Discuss language and discourse measures that can contribute to understanding student learning.
2. Discussion on possible reasoning behind the observed trends (e.g. the readability level of the information giver is low (i.e. discourse complexity is high) when compared to the information seeker and other user roles, the level of information embeddedness (number of clauses) is high within the information giver compared to the remaining classes).
3. Discussions on understanding the correlations between external factors (e.g. course structure, learners' demographic) and learner's role (e.g., information seeker, information giver) transformations.
4. Discussions on how existing learning frameworks (e.g. ICAP framework) associate with learner roles.

7. REFERENCES

[1] Anderson, T., 2004. Towards a theory of online learning. *Theory and practice of online learning 2*, 109-119.

[2] Arguello, J. and Shaffer, K., 2015. Predicting speech acts in MOOC forum posts. In *Ninth International AAAI Conference on Web and Social Media*.

[3] Crossley, S.A., Greenfield, J., and McNamara, D.S., 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly 42*, 3, 475-493.

[4] Daphne Koller, Ng, A., Do, C., and Chen, Z., 2013. Retention and Intention in Massive Open Online Courses: In Depth. *Educ. Rev. 48*, 3, 62-63.

[5] Dowell, N.M., Brooks, C., Kovanović, V., Joksimović, S., and Gašević, D., 2017. The changing patterns of MOOC discourse. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, 283-286.

[6] Hecking, T., Chounta, I.-A., and Hoppe, H.U., 2016. Investigating social and semantic user roles in MOOC discussion forums. In *Proceedings of the sixth international conference on learning analytics & knowledge*, 198-207.

[7] Lundberg, J., Castillo-Merino, D., and Dahmani, M., 2008. Do online students perform better than face-to-face students? Reflections and a short review of some empirical findings. *RUSC. Universities and Knowledge Society Journal 5*, 1, 35-44.

[8] Nguyen, D. and Rosé, C.P., 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media* Association for Computational Linguistics, 76-85.

[9] Searle, J.R., 1976. A classification of illocutionary acts. *Language in society 5*, 1, 1-23.