Challenges of Identifying Second Language English Speakers in MOOCs

Ismail Duru*, Ayse Saliha Sunar**, Gulustan Dogan*, and Su White**

(*)Yildiz Technical University, Istanbul, Turkey
(**)The University of Southampton, UK
{iduru,gulustan}@yildiz.edu.tr
{ass1a12,saw}@ecs.soton.ac.uk

Abstract. In this study, we aim to analyse English as a Second Language (ESL) and English as a First Language (EFL) MOOC participants' engagements in a MOOC. We aim to find out key points which directly effect learners' dropout and performance in MOOCs. We worked on a FutureLearn data which is provided by the University of Southampton. The course is Understanding Language: Learning and Teaching MOOC that was run between 2016-04-04 and 2016-05-02 is chosen for the analysis. According to the results, it is very challenging to identify who is a second language English speaker by using their location information. One of the important findings is that first language English speakers wrote longer comments. In order to identify strategies for ESL MOOC participants, which is one of the ultimate goal of our research, there is a need for much deeper analyses.

Keywords: Second language English speakers, MOOC, Online learning, Dropout, Predictive models, Learner behaviour

1 Introduction

Online platforms which offer online courses with free registration to any learners who would like to participate, become one of the trend implementations in technology enhanced learning and are investigated from many different perspectives [3, 7, 10, 12]. These courses are adverted as Massive Open Online Course (MOOC) and many institutions from all around the world have attempted to build MOOC platforms.

Even though these courses attract millions of learners from numbers of different countries, the mainstream MOOC providers are based on English speaking countries such as the US and the UK[1]. According to the 2015 statistics that *class-central* published, 75% of the MOOCs have been offered in English. This rate was 80% in 2014. The website published different kind of review analysis for 2016. They identified that around 25% of the new MOOC learners were attracted by the local providers that offer MOOCs in languages other than English[2].

[1] https://www.class-central.com/report/moocs-2015-stats
[2] https://www.class-central.com/report/moocs-stats-and-trends-2016

Additionally, many researchers have identified that one of the reasons for the low completion rate, which is one of the main concerns in MOOCs, is their one-size-fits-all model [1, 5]. MOOC content does not change for individual students according to their needs such as language, difficulties, and learning approaches.

Some locally launched MOOC platforms provide courses in local languages for the targeted audience. For example, EMMA, a European MOOC platform, provides MOOCs in multiple European languages[3]. Japanese MOOC platform (JMOOC) stated on their website that since many Japanese are struggling with studying in English, they launched JMOOC to serve the Japanese speaking communities[4].

Nevertheless, ESL speakers continue to pay attention to MOOCs that are offered in English. In order to help these learners engage with the platform, some personalised services are provided. For example, Coursera asks their attendees to voluntarily contribute to translation of course content [5]. Herewith, learners may benefit from the course at a higher level.

However, to the best of our knowledge, there is not much studies available to investigate possible differences between engagements of first and second English language speakers in MOOCs. Therefore, our conducted research sought to address this problem. In this paper, we specifically focus on the identifying second language English speakers and their common engagement patterns. To follow up this study, we will research on i) predicting participants' future participation and certificate earn using identified engagement patterns ii) recommending specific identifiable strategies for the ESL speakers when working in a MOOC.

2 Related Works

Barak et al. [2] suggested that learners who have high motivation are likely to complete the course. Additionally, learners who study in their first language have higher confidence to finish the course and this increase their motivation level.

Eriksson et al. [4] put forward in their qualitative study investigating the reason of learners' dropouts that some learners had struggled with understanding the spoken language in the video and occasionally the instructors' accent. The MOOC participants also stated that subtitles in English were helpful.

There are numbers of studies that investigate how learners engage with MOOCs to identify and classify patterns of learner behaviours (i.e., [6, 8]). Researchers use leverage on statistical and learning analytics methods and machine learning techniques to classify learners' based on their behaviours predominantly based on activity logs data and click-stream data in courses. For example, Milligan et al. [9] conducted interviews and classified learners based on their statements and course activities. Gillani et al. [6] use participants' statements reflected by their comments in discussion threads for classification.

[3] https://platform.europeanmoocs.eu
[4] http://www.jmooc.jp/en/about
[5] http://www.coursera.community/#gtc

Some researchers focus on different factors for identifying and predicting level of engagement and course completions, which is one of the common objectives in MOOC research. For example, Kizilcec et al. [8] mainly consider timely assessment submissions and identified learners' behaviour patterns as *auditing, behind, on track,* and *out*. Then, they group learners based on their level of engagement as *auditing, completing, disengaging,* and *sampling*

However, to the best of our knowledge, there is not much many studies specifically emphasising on identifying ESL speakers and their engagement performance in MOOCs. Uchidiuno et al. [11] used browser language preferences of participants in a MOOC and analysed their video interactions. The authors found that using browser language preference is helpful to more accurately identify ESL speakers.

Our research also aims to contribute to the research in this area by identifying needs and behaviours of ESL participants in MOOCs. The finding of our research ultimately may help ESL speakers, and the instructors who may use MOOCs in foreign language as a material on their blended campus education.

3 Methodology

In order to identify second language English speakers, we first used the data about learners' location, which is generated from the pre-course survey on the platform. Then, we used participants' comments in the discussions to gather more information about their location and first language.

In this paper, we sought to do some investigation to find the best and most accurate method to group learners based on their first languages to analyse their behavior and predict their future performance. Further in the research, we will use machine learning techniques and natural language processing to automatically identify ESL speakers.

3.1 Datasets

We have used the forth run of the *Understanding Language: Learning and Teaching* MOOC (UL-MOOC) that was run between 2016-04-04 and 2016-05-02. We have chosen this course as a start since it has attracted many international English language teachers around the world. While dataset includes 25598 user records, only 3306 of them had location information.

Data files that we used in this study are as follows.

- Enrolments: Includes demographic information of participants, enroll and unenroll time, and purchased statement certificate information.
- Step Activity: *Step* is used for each learning unit in the course. Each week is consisted of numbers of steps. This dataset includes the visit and completion time information for each learner.
- Comments: Includes comment priorities related to course structure, comment text, commenting learner and time.

4 Results of Analysis

In order to identify ESL and EFL speakers in the MOOC, we took three different approach.

1. Divided into two groups as ESL and EFL speakers by using only their location.
2. Divided into three as primary ESL, not primary ESL, and EFL speakers by using only their location.
3. Updated the groups based on participants' comments.

4.1 Implementation of the First Stage

In this step, ESL participants are identified based on the country information only. In order to group countries by the country's official languages, we have utilised a Wikipedia article as a reference[6]. According to the languages, the groups are as follows: i) English as a first language learners (EFL), ii) English as a second language learners (ESL), and iii) no country information presented in the database.

The results showed us that there is no statistically significant difference amongst the course performance of learners in each group. The reason could be the inefficient way of grouping learners by the countries' official languages. Therefore, we have conducted the second step for better categorisation.

4.2 Implementation of the Second Stage

In the first stage, India and the United Kingdom are in the same group. Even though English is very commonly used language in India, it is not the primary language. To have better understanding of the demographics of subgroups to identify their behaviours, we now divided the countries into: i) English as official and primary language and ii) English as official but not primary language, and iii) English as second language. Here, we have again used the same Wikipedia article.

According to the country information of learners, 18% of participants join from a country of which English is an official and primary language, 11% of them join from a country of which English is an official but not primary language, and lastly 71% of them are English as a second language learners.

Fig. 1 shows that what percentage of participants in each group completed at least one learning step in the course. There is no big difference between one-time-show-up ratio amongst the participant in each group.

[6] https://en.wikipedia.org/wiki/List_of_territorial_entities_where_English_is_an_official_language

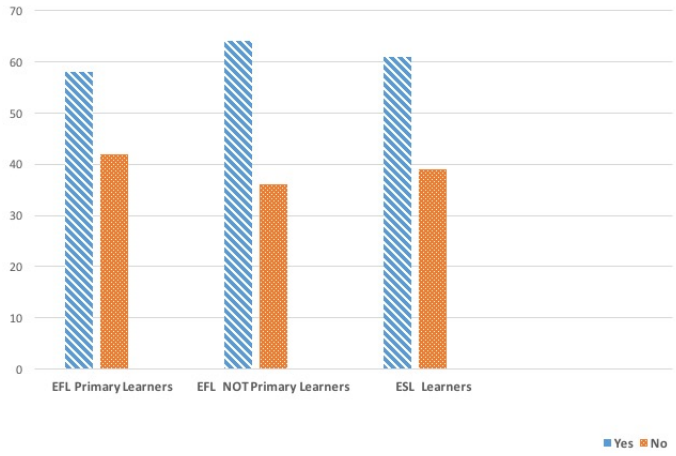

Fig. 1. Ratio of Learners in Each Group who completed at least one step

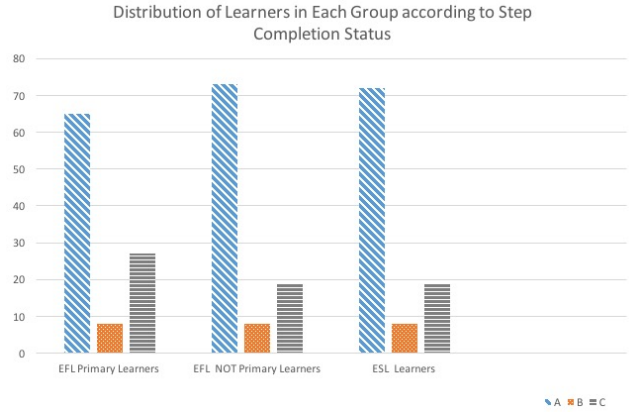

Fig. 2. Distribution of Learners in Each Group according to Step Completion Status

Fig. 2 presents the ratio of learners based on their course completion status. In the figure, A indicates a completion of less than half of the steps, B indicates a completion of more than half of the steps but less than 80% of the steps, and C indicates a completion of more than 80% of the total numbers of steps. Learners in the group of English as a official and primary language outperform in completing more than 80% of the steps.

Fig. 3 shows the percentage of learners in each group who contributed to the discussions with at least one comment. However, there is no big difference between groups in terms of presence in discussion forums. There is a need for a deeper investigation in forums. Therefore, we have checked the length of learners' comments in the discussion forums.

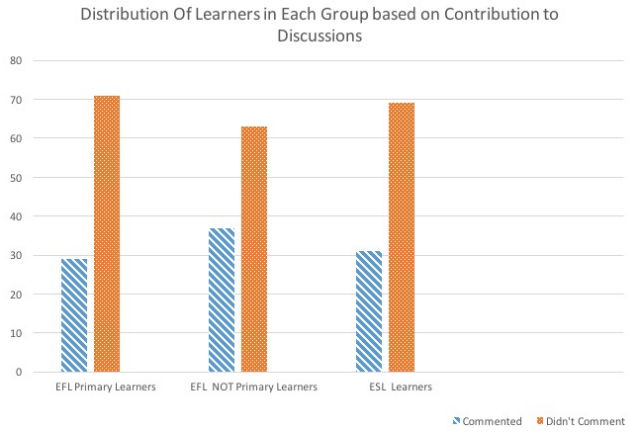

Fig. 3. Distribution Of Learners in Each Group based on Contribution to Discussions

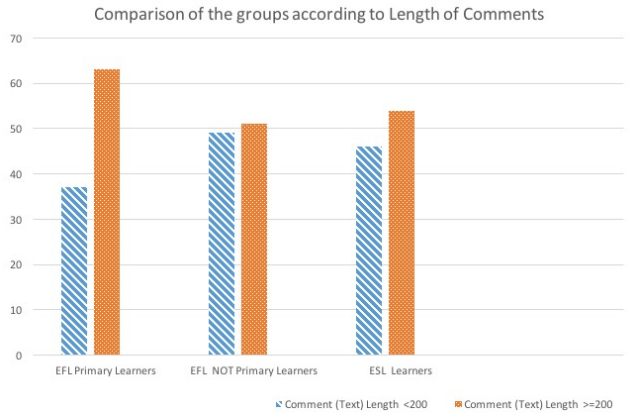

Fig. 4. Comparison of the groups according to Length of Comments

Fig. 4 compares the length of comments written by learners in each group. The figure shows that higher percentage of the learners in the group English as a official and primary language wrote comments that contains at least 200 characters.

4.3 Implementation of the Third Stage

In the previous section, the results showed us there is not much difference between groups, especially between the learners in the English as official but not primary and the ESL learners. The reason could still be the inefficient way of grouping learners.

Indeed, when we group the learners according to where they live (which is the country information in the dataset), we sometimes miss the learners who speak a different language than the official/primary language of the country that they live in. Therefore, in the third step, we attempted to investigate comments that were posted to discussion forums.

The course instructors asked the participants how they use language in their daily life in the Step 1.5 of the first week. In the associated discussion forum to Step 1.5, participants gave information about where they live, what their first language is, and how fluent they are in English.

We updated the groups according to the additional information we acquired from learners' comments in Step 1.5.

We made the same analysis for the new updated groups. If we see different distributions than what we had in the first two stages of the analysis, we could say that the results for each group is mostly similar is not because this is the situation but there is a need for more precise grouping method. Number of the users gave information about their first language and had country information were 694. According to the comments on Step 1.5, we realised that 68 of the records were grouped in a wrong cluster in the previous stage, which is nearly 10% of the 694 records. Although some of other learners also commented in Step 1.5 they had no information about their location in the database. We have updated the groups according to the information we have acquired from the discussion thread. This stage indicates that a deeper and systematic analysis in learners' comments can help us to improve our grouping model with the aid of machine learning.

Fig. 5 shows distribution of step completion status for participants who commented in Step 1.5 and have country information. According to Fig. 5, there is a huge increase in step completion rate.

Fig. 6 shows distribution of the length of comments for learners who commented in Step 1.5 and have country information. The results indicates bigger difference than we have seen in the first two stages of the analysis. According to the new results, 80% of learners who commented in Step 1.5 and speak English as first language which is the primary and official language of the country they live in, wrote comments that are longer than 200 characters.

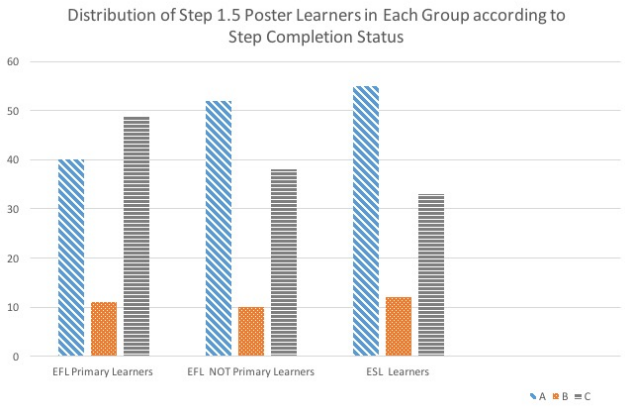

Fig. 5. Distribution of Step 1.5 Poster Learners in Each Updated Group according to Step Completion Status

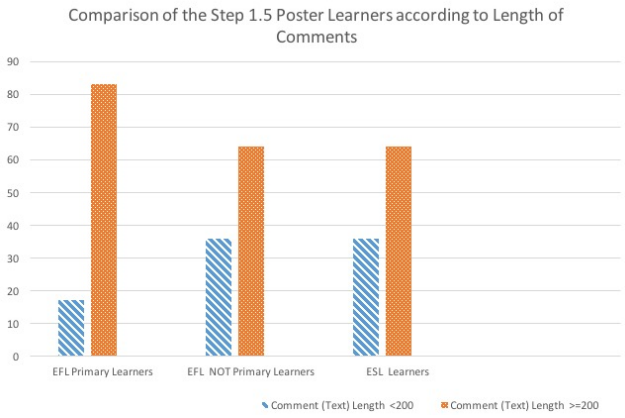

Fig. 6. Comparison of the Step 1.5 Poster Learners in Each Updated Group according to Length of Comments

5 Discussion and Future Work

In this study we sought to i) identify English as first and second language participants in a MOOC, and ii) investigate and interpret differences between their overall course engagement.

Firstly, we grouped learners based on their country information as "English as a Official and Primary Language", "English as a Official but Not Primary Language", and "English as a Second Language" group.

Secondly, to have more accurate group separation, we investigated learners comment in the discussion forum of Step 1.5, where learners introduce themselves with country and language information. With this information, we have updated our groups.

However, the manual identification of ESL participants from their comments is time consuming and not reliable for implementing at massive numbers of learners in more than one MOOC. In the process of identifying ESL learners, the next step is to use a machine learning technique (a suitable technique has not determined yet) and natural language processing to accurately identify EFL and ESL participants in a MOOC. Additionally, we will take into account of fluency level of ESL learners in our future research.

We plan to do further data analysis to understand the characteristics of our data. In each group we will do outlier detection and will decide about how uniformly the groups behave. Then, we will run some correlation models to find out the highly correlated fields to determine a feature set and a prediction model. We will also analyse differences between social engagements of the learner groups. As a result of our studies, we aim to build a prediction model for the dropouts of the ESL learners.

We believe that our results can be a new perspective for MOOC personalisation. As we do not hold the source code for popular MOOC platforms, we cannot make direct changes to the platforms. However, these identification methods can be used by MOOC providers for identification of ESL students. An ESL student who uses any MOOC platform such as FutureLearn would be more engaged if they are given a more personalised experience.

Acknowledgments. This work has been done under the project numbered 2016-04-01-DOP05 in Yildiz Technical University (YTU). The dataset used in this paper is provided by the University of Southampton for the ethically approved collaborative study (ID: 23593). The authors would like to thank Prof. Dr. Banu Diri from YTU for her help in improving groups based on learner comments. Authors also want to thank Hatice Ata and Ali Demir who are undergraduate students in YTU for their help in analyses.

References

1. A. Bakki, L. Oubahssi, C. Cherkaoui, and S. George. Motivation and engagement in moocs: How to increase learning motivation by adapting pedagogical scenarios? In *Design for Teaching and Learning in a Networked World*, pages 556–559. Springer, 2015.
2. M. Barak, A. Watted, and H. Haick. Motivation to learn in massive open online courses: Examining aspects of language and social engagement. *Computers & Education*, 94:49–60, March 2016.
3. F. Brouns and O. Firssova. The role of learning design and learning analytics in MOOCs. In *Conference Proceedings of the 9th EDEN Research Workshop*, Oldenburg, Germany, October 2016. EDEN.

4. T. Eriksson, T. Adawi, and C. Stöhr. "time is the bottleneck": a qualitative study exploring why learners drop out of MOOCs. *Journal of Computing in Higher Education*, pages 1–14, November 2016.
5. H. A. Fasihuddin, G. D. Skinner, and R. I. Athauda. Boosting the opportunities of open learning (moocs) through learning theories. *GSTF Journal on Computing (JoC)*, 3(3):112, 2013.
6. N. Gillani, R. Eynon, M. Osborne, I. Hjorth, and S. Roberts. Communication communities in MOOCs. *arXiv preprint arXiv:1403.4640*, 2014.
7. L. Guàrdia, M. Maina, and A. Sangrà. MOOC design principles: A pedagogical approach from the learner's perspective. *eLearning Papers*, (33), 2013.
8. R. F. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *The 3rd international conference on learning analytics and knowledge*, Leuven, Belgium, April 2013. ACM.
9. C. Milligan, A. Littlejohn, and A. Margaryan. Patterns of engagement in connectivist MOOCs. *Journal of Online Learning and Teaching*, 9(2):149, July 2013.
10. A. Sunar, S. White, N. Abdullah, and H. Davis. How learners' interactions sustain engagement: a MOOC case study. *IEEE Transactions on Learning Technologies*, December 2016.
11. J. Uchidiuno, A. Ogan, K. R. Koedinger, E. Yarzebinski, and J. Hammer. Browser language preferences as a metric for identifying esl speakers in moocs. In *Proceedings of the Third ACM Conference on Learning@ Scale*, pages 277–280, 2016.
12. S. White and S. White. Learning designers in the 'third space': The socio-technical construction of moocs and their relationship to educator and learning designer roles in he. *Journal of Interactive Media in Education*, 2016(1), November 2016.