FORUM ANALYSIS TO AID IN MOOC RECOMMENDATION

Project ID: 19-089

Research Thesis

Hilmy S.B.M.

B.Sc. (Hons) Degree in Information Technology

Department of Computer Science & Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

April 2019

FORUM ANALYSIS TO AID IN MOOC RECOMMENDATION

Project ID: 19-089

Research Thesis

B.Sc. (Hons) Degree in Information Technology

Department of Computer Science & Software Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

April 2019

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as article or books).

Student Name	Student ID	Signature
Hilmy S.B.M.	IT16005372	

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

Supervisor: Mr. Nuwan Kodagoda

Co-supervisor: Ms. Kushnara Suriyawansa

Signature: Date:

Signature of the supervisor: Date:

Signature of the co-supervisor: Date:

ABSTRACT

Massive Open Online Courses (MOOCs) are a popular learning medium in the current day and age. Its popularity has been increasing steadily over the years and now are helping people all around the world learn at an unprecedented scale. When it comes to learning anything a proven to be effective way of doing this is by discussing the topic with peers. This trait of learning is facilitated by the MOOCs and their platforms in the form of 'Forums'. When it comes to learning through this method or enhancing the learning effectiveness, some people prefer it more than others. In fact, research has shown that as a result of their preference towards learning through discussion they learn faster when engaging in subject content related discourse. Forums in general contain a large amount of information which can be used to produce useful insights to the students of the course. This thesis presents the analysis of MOOC forums to generate useful insights which can in turn help students choose the course they are going to follow. Raw forum data is gathered from different MOOC platforms with the use of web crawling techniques. Various analytical methods are used to get useful information from the retrieved data which includes the Natural Language Processing technique (NLP) sentiment analysis.

Keywords: MOOCs, Forum Analysis, Data Mining, Sentiment Analysis

ACKNOWLEDGEMENT

This academic thesis and work described in it were conducted as a part of 4th year research module (Comprehensive Design Analysis Project) under the degree program Bsc (Hons) in Software Engineering in Sri Lanka Institute of Information Technology. The fruition of this research and the associated software project are results of dedication of the 4 group members and the immense support given by a supervising panel.

Thus, we would like to express our gratitude towards Dr. Nunan Kodagoda and Ms. Kushanara Suriyawansa, who are our supervisors as well as lecturers of Sri Lanka Institute of Information Technology. Completion of this academic work would not have been possible without their guidance and insight. We are extremely grateful to Dr. Malitha Wijesundara and Dr. Dasuni Nawinna, who reviewed and approved our research topic and provided valuable suggestions. Furthermore, we wish to express our gratitude to Mr. Jayantha Amaraarachchi, the Head of SLIIT Centre for Research who provided us with immense knowledge and guidance throughout.

Also, we would like to express our gratitude to all our colleagues and companions who helped us in testing and in various other ways to make our research come to fruition. At the same time, we would like to thank everyone who might not have been mentioned but has given their support and encouragement to us.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

MOOC	Massive Open Online Course
NLP	Natural Language Processing
ML	Machine Learning
VADER	Valence Aware Dictionary and sEntiment Reasoner

1 INTRODUCTION

1.1 Background Literature

MOOCs are online courses which usually reside on platforms like Coursera, FutureLearn, Edx, etc. Though MOOCs are a relatively new phenomenon they have rapidly gained in popularity by learners worldwide. MOOCs gives access to high quality educational content to any user who has an internet connection. According to a study conducted by Class Central 101 million people were following MOOCs in 2018 [1]. A study conducted by Hanan et al [2] show that a high number of learners drop off while having followed the course only halfway and shows the possible causes for the drop-out rate. The paper suggests different approaches to tackling the problem of learner retention one of which is enhancing 'student to student' and 'student to instructor' interactions.

This thesis is about improving the existing recommendation system in the Mooc Rec v2 system by taking the forums of the MOOCs into account when recommending. Forums are the place where online interaction takes place, and helping the users find MOOCs where the forums are the most active will help improve user retention. The study done by C. Kent et al [3] also shows that interactivity in online discussions improve the learning outcomes of students.

iRobot is an intelligent web crawler which can identify web forums with minimal initial data and figure out how to traverse the forum [4]. However, iRobot is a general-purpose web crawler which is mainly designed to comb through large numbers of web sites with minimal modification and extracts only a small amount of information from each web page. The approach taken in this study would not suit the purpose of this thesis, which is to gather a large amount of structured information from a specific platform (web site).

The web crawler described in [5] had aimed at gathering data using regular expressions which is a template-based processing method. As a result, the information acquired is structured which makes it easier to process them later, but the disadvantage to this method is that the regular expressions must be explicitly written for the data that the system is attempting to collect. So, the system could miss out on potential information if it is not designed to do so because regular expressions collect only specific data points.

A.Pak and P.Paraoubek have presented a general sentiment analyzing model which was trained using 300,000 twitter posts [6]. The results of the research paper show that the model was able to identify user sentiments with reasonable accuracy. A multinomial Naïve Bayes classifier had been chosen to use the model data to make the classifications. A significant limitation of this classifier is that it is limited to the English language. MOOCs exist in a variety of languages and the sentiment analyzer must be capable of handling those languages.

Sentiment analysis has been done in a much different approach by B. Pang et al [7]. Before the data is used to train or classify text, the system proposed compresses and removes any unnecessary subject domain text while retaining polarity information. Basically, it compresses the data before processing, as a result the performance was increased. In addition to that, the process increases the accuracy because the removal of unnecessary data makes the data cleaner (reduced noise) and has a lower chance of being misidentified by the Naïve Bayes classifier.

Though previous research has shown positive results in the field of web crawling and sentiment analysis, they were not regarding the context of discussion forums relating to MOOCs. Therefore, research is needed to determine to which extent information (ratings derived from sentiments, other forum thread metadata) contained within forums can be used to improve MOOC recommendations.

1.2 Research Gap

MoocRec V1 introduced novel and useful features that did not exist in any MOOC recommendation software previously [8]. Though MoocRec V1 offered benefits, a literature review was conducted, and it identified certain characteristics of the tool which could be improved upon and new features that would improve the software's usefulness. One of the identified useful features was 'Forum Analysis to Aid in MOOC Recommendation'. This feature does not currently exist in any MOOC platform but would potentially benefit student who are searching for MOOCs.

1.3 Research Problem

This component has essentially two phases in its process which is forum data gathering and forum data analysis. The forum data resides within the MOOC platforms and would need to be retrieved from them. The forum data needed for the analysis phase cannot be retrieved from the MOOC platforms using their APIs as they do not currently support providing that data to third-parties. A simplistic web crawler which retrieves the HTML code and extracts the relevant information from it is insufficient, because most required information is locked behind a login. For this reason, the web crawler should be able to perform authentication to access secured resources.

By observing the authentication mechanisms of the MOOC platforms, it was found that the process is complex and infeasible to be implemented in a simplistic web crawler, so retrieving the web pages as code and scraping data from it would not be sufficient. This is because there are many security mechanisms built into the web pages such as CSRF tokens and the web crawler would have to implement those mechanisms into itself, which is not feasible. Several alternatives approaches are needed to be tested to find a reasonably efficient method of gathering forum data.

Forum data changes of MOOCs change over time; therefore, the whole process has to be repeated after a certain period of time. But for this process to repeatedly occur it has to be efficient and have good performance. A few wasted milliseconds in each MOOC can result in many extra hours being added into the processing loop.

Another problem is that the level of detail and the types of details the forum data contained within these platforms differ from each other. This creates a problem when calculating the ratings because the resulting ratings of each course is arrived at using varying amounts of data and therefore cannot be compared fairly with each other. The process of calculating the ratings should compensate for the differences in available data. When the 'course rating' is calculated the already existing course rating which is available in the MOOC platform should be factored in to make use of that information.

1.4 Research Objectives

The main focus of the forum analysis component is to improve the recommendations of MOOCs using insights generated from data residing within forums, which includes the threads and other meta-data. This main objective can be further divided into three, they are:

1. Gathering raw forum data from the three different MOOC platforms Coursera, Edx, & FutureLearn in an efficient manner.

2. Analyze the forum data to calculate the 'Course Rating' and the 'Forum Activity Rating'.

3. Calculating the ratings in such a way that it compensates for the differences in available data for each MOOC.

2 METHODOLOGY

2.1 Methodology

The forum analysis process requires a significant amount of data to be able to calculate the ratings. The larger the amount of data the more accurate the ratings tend to be as outliers are balanced out.

To satisfy this, three web crawlers were implemented to gather data from the MOOC platforms Coursera, Edx and FutureLearn [9]–[11]. The Coursera and FutureLearn crawlers gather data by simulating a web browser and extracting data from the rendered web pages. The Edx crawler uses a hybrid approach as it simulates a web browser to authenticate and navigation purposes but collects the HTTP responses

while they are on their way to the browser. The Edx crawler is capable of gathering more in-depth information than its counterparts due to the fact that it extracts JSON data which are received directly from the server.

Once the data gathering processes have concluded the analysis process will begin. The result of the analysis process is two rating scores, the overall course rating score and the forum activity score. The overall course rating is a measure of how useful the learner the course is while the forum activity score is a measure of how active a forum is. The high-level diagram of the forum analysis component is shown in Figure 2.1.

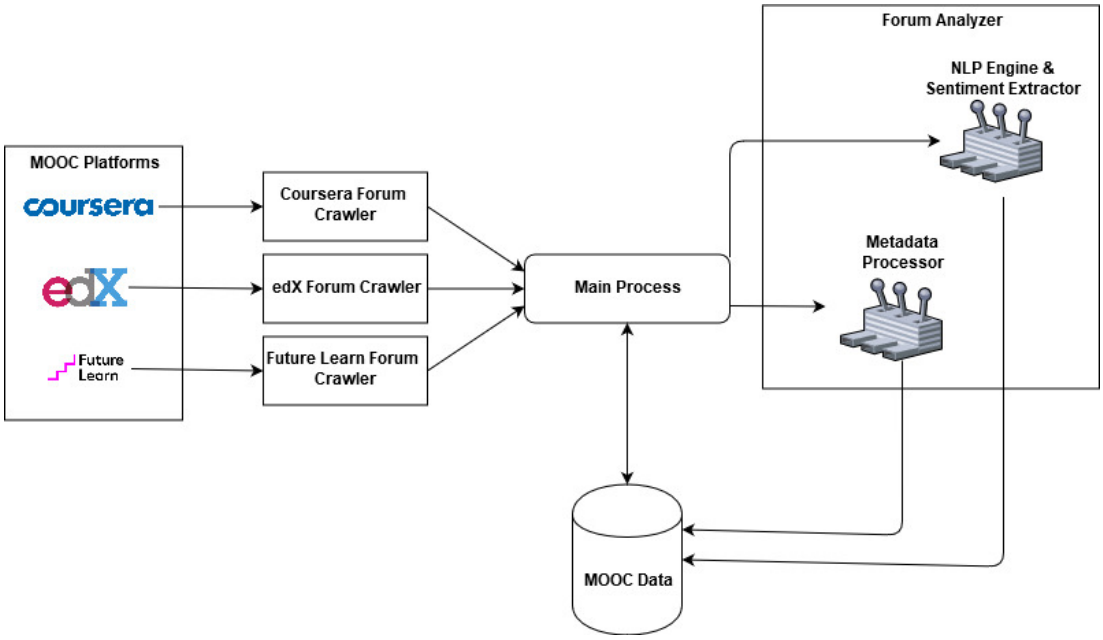

Figure 2.1: System Architecture

Online discussion forums regarding MOOCs generally fall into two categories; questions or discussions. Questions are threads in which discussions about the course content takes place. On the other hand, discussion-type threads are regarding the MOOC itself i.e. how well the teacher explains the topic, the language fluency, whether the topic of the subject is covered, etc. Complaints usually fall into this category and will be detected by the sentiment scores of the thread text data as it will show up as having relatively lower scores. Sentiment of text refer to how positive or negative the piece of text is.

The course rating is calculated based on the course rating available on the MOOC platform the course was extracted from and the sentiment scores of discussion-type threads as shown in the equation below. Equal weighting is given to the calculated course rating and the already existing course rating on the platform.

$$C = (0.5 \times P) + (0.5 \times S)$$

Where,

C = Calculated Course Rating (Range – Between 0 and 5)

P = The rating the course has on its platform (Range – Between 0 and 5)

S = The sentiment score calculated based on the discussions (Range – Between 0 and 5)

The forum rating which is separate from the course rating is calculated based on the statistics derived from the metadata of the forum threads. The correlation of the attributes is determined manually based on whether they positively affect the forum rating or negatively affect it, for example the higher the no. of unique learners the better it is, therefore it is considered a positive factor. Neutral attributes are not considered when the forum rating is calculated as its effect on the forum rating is not clear enough.

The metadata attributes used to calculate the forum rating:

Table 2.1: Metadata Attributes

Attributes	Positive or Negative Correlation with Forum Activity
No. of threads per month	Positive
Total no. of threads	Positive
No. of replied threads	Positive
No. of threads not replied to	Negative
No. of question-type threads	Neutral
No. of discussion-type threads	Neutral
No. of unique learners	Positive
Total no. of posts	Positive
Avg. no. of posts per thread	Positive
No. of days since thread was last active	Negative

$$FR^{course} = \frac{Product\ of\ Positively\ Affecting\ Attributes}{Product\ of\ Negatively\ Affecting\ Attributes}$$

Where,

Positively Affecting Values = Attributes which have a positive correlation with forum activity

Negatively Affecting Values = Attributes which have a negative correlation with forum activity

FR^{course} = Forum Rating of the Course

Since the intermediate value has a very large variance, it needs to be normalized and put into the same scale of values. Normalization is done in two steps, the first step applies the new values based on the following equation, this puts all the values in a similar scale.

$$FR^{course}_{normalized} = \sqrt[4]{FR^{course}}$$

Where,

FR^{course} = Intermediate Forum Rating of the course calculated in the previous step

$FR^{course}_{normalized}$ = The new normalized forum rating of the course

In the second step of normalization the values are strictly converted into values between 0 and 5 using the following equation.

$$FR^{course}_{final} = \frac{FR^{course}_{normalized}}{\max\left(FR_{normalized}\right)} \times 5$$

Where,

$FR^{course}_{normalized}$ = The normalized forum rating of the course

$FR_{normalized}$ = The set of normalized forum ratings of all courses

FR^{course}_{final} = The final fully normalized forum rating of the course (Range – Between 0 and 5)

2.2 Testing & Implementation

The forum analysis was implemented as described in the methodology (Section 2.1) and then it was measured for how well it performs. Performance measurements was performed on three selected MOOCs of varying popularity and three readings were taken which were then averaged to get the actual measurement.

Initially the three web crawlers are used to gather raw forum data from the MOOC platforms Coursera, Edx, and FutureLearn.

The gathered raw data is then processed by the forum analyzing subcomponent which performs meta-data analysis and sentiment analysis. This produces two rating scores, the 'Course Rating' and the 'Forum Activity Rating'. These two ratings will be used to improve the current MOOC recommendation system of MoocRec V1.

3 RESULTS & DISCUSSION

3.1 Results

The performance of the three web crawlers were measured in terms of speed and efficiency. The efficiency of the crawlers in retrieving data from courses is shown in Table 2.1. The speed of the web crawlers is shown in Table 2.2. Three courses from the three platforms were selected to be used to the take performance readings.

With regards to Table 2.1, the 'actual no. of threads' data points were available on the respective platforms. The 'extracted no. of threads' data were tracked by the web crawlers itself.

The hardware specifications of the computer used to test the system and other relevant details:

- Processor: Intel Core i7-6700HQ 2.60 Gigahertz
- RAM: 16GB DDR4
- Avg. Network Speeds
    - Avg. Download Speed: 64.93 Mbps
    - Avg. Upload Speed: 47.61 Mbps
- Storage Medium: SSD

The data shown can be considered accurate as it deals with no. of threads which is a discrete value. The courses were selected as to maximize the variance of the no. of threads across the different platforms.

Table 3.1: Web Crawler Efficiency

Platform	Course	Actual No. of Threads	Extracted No. of Threads	Effectiveness (%)
FutureLearn	Introducing Robotics	242	242	100%
FutureLearn	An Introduction to Cryptography	45	45	100%
FutureLearn	Python in High Performance Computing	170	169	99.4%
Edx	Python for Data Science	1,971	1,950	98.9%
Edx	Software Engineering: Introduction	1,034	1,030	99.6%
Edx	C Programming: Advanced Data Types	646	640	99.1%
Coursera	Python for Everybody	41,230	41,185	99.8%
Coursera	Neural Networks and Deep Learning	11,420	11,420	100%
Coursera	Google Cloud Platform Fundamentals: Core Infrastructure	2,685	2,685	100%

With regards to Table 2.2, the elapsed time was tracked by the crawlers themselves. The data shown is an average of five separate executions of the task. The variance was the time taken is 3-6%. This time information includes the amount of time taken to save the extracted information to the database which was hosted locally to minimize the database handling performance impact.

Table 3.2: Web Crawler Performance Data

Platform	Course	Time Taken to Extract Data (seconds)	Variance (σ^2)
FutureLearn	Introducing Robotics	77 (1 min. 17 sec)	2.30
FutureLearn	An Introduction to Cryptography	34	1.30
FutureLearn	Python in High Performance Computing	56	2.10

Edx	Python for Data Science	498 (8 min. 18 sec.)	22.4
Edx	Software Engineering: Introduction	283 (4 min. 43 sec.)	11.2
Edx	C Programming: Advanced Data Types	144 (2 min. 24 sec.)	5.50
Coursera	Python for Everybody	1162 (19 min. 22 sec.)	67.7
Coursera	Neural Networks and Deep Learning	355	25.6
Coursera	Google Cloud Platform Fundamentals: Core Infrastructure	173	12.4

Average time taken to gather data from a course = 309.1 seconds (5 min. 9.1 sec.)

The system generated forum ratings were evaluated for accuracy by asking a set of students of users to go through the course and forums manually and rating the forums out of 10 (scaled down to 5 for comparison). The users were asked to first go through all the courses to get a rough idea of how the activity vary among the courses and then rate each one of them by going through them a second time.

User information:

- Age: 20-25 years
- Educational Level: Undergraduate
- Number of testers: 10
- Accuracy of Information: ± 0.5

Table 3.3:Forum Rating Accuracy Evaluation

Platform	**Course**	**Forum Rating**		**Difference (Error)**	
		System Generated (±0.1)	Avg. User Rating (±0.5)	Value	Percentage

FutureLearn	Introducing Robotics	2.9	2.5	0.4	8%
FutureLearn	An Introduction to Cryptography	1.3	2.5	1.2	24%
FutureLearn	Python in High Performance Computing	1.9	2.5	0.6	12%
Edx	Python for Data Science	3.7	4.0	0.3	6%
Edx	Software Engineering: Introduction	2.2	2.0	0.2	4%
Edx	C Programming: Advanced Data Types	1.7	2.0	0.3	6%
Coursera	Python for Everybody	4.8	4.5	0.3	2%
Coursera	Neural Networks and Deep Learning	3.8	4.0	0.2	4%
Coursera	Google Cloud Platform Fundamentals: Core Infrastructure	2.7	3.0	0.3	6%

3.2 Research Findings

Average error of the proposed system = 8.0%

The average error of the proposed system is 8.0% which means the system has an average accuracy rating of 92%. This level of accuracy at determining the forum rating of the courses is good enough to give the user an enhanced experience. The Thread Count-vs-Error information is plotted in the graph shown in Figure 3.1 (X-axis: Thread Count, Y-axis: Error).

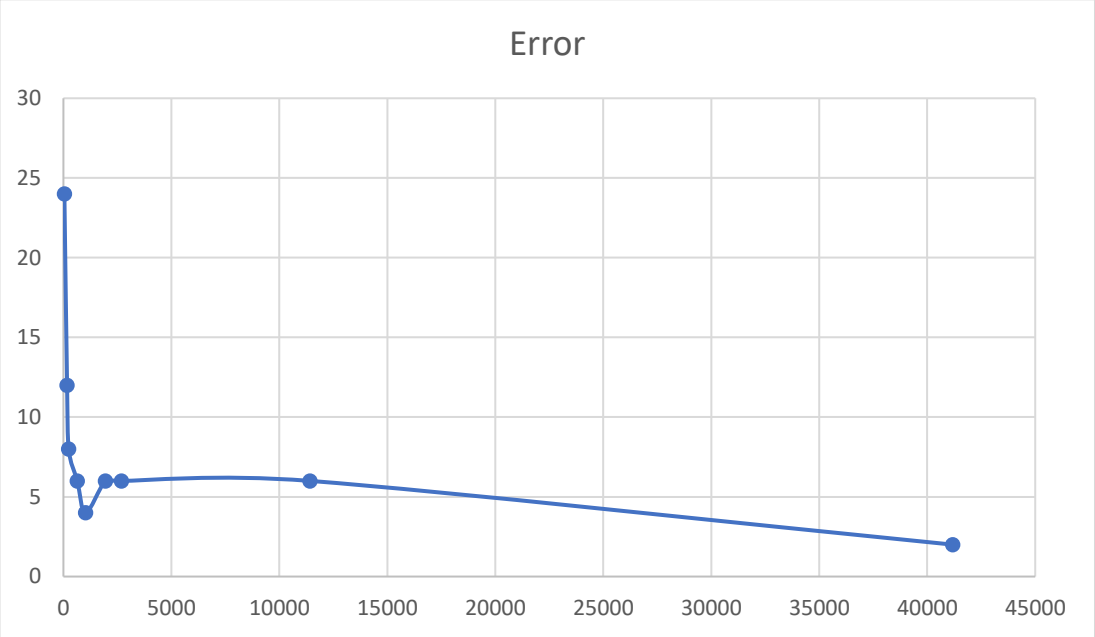

Figure 3.1: Thread Count vs Error Graph

The measured web crawler performance times indicate that the average time taken to gather data from a single MOOC is 309.1 seconds (5 min. 9.1 sec.). Since the courses used for the tests were carefully selected so that they represent the entire population of courses, the average time taken is an accurate enough value for the purposes of the research.

The total no. of available MOOCs on all platforms (2018) [1]= 11, 400

Assumption: Equal amounts of courses are available on the 3 MOOC platforms (Coursera, Edx, and FutureLearn).

$$ETD = \frac{TC \times ATD}{3}$$

Where,

ETD = The estimated time duration to gather data of all MOOCs

TC=Total no. of existing MOOCs (11,400)

ATD = Avg. Time Duration to gather data from a single MOOC (309.1 seconds)

Note: The product of the values is divided by a factor of 3 because the 3 web crawlers can run in parallel

When solved,

ETD = 1,174,580 seconds (≅13 days)

Which means it takes 13 days to gather all the forums from all the existing MOOCs.

3.3 Discussion

The measured web crawler performance times indicate that 13 days will be required for the data gathering process. This is not issue because forum data is not very volatile and does not change at rapidly. The efficiency of the web crawler is close to 100% which means it extracts almost all threads with minimal loss. The design of the web crawler is sufficiently dependable to gather forum data reliably and make it available for the forum analysis process.

The Figure 3.1 shows that the error reduces as the thread count rises. The reason for this progressively increasing accuracy is probably the additional data compensating for the inherent noise and outliers which usually exist in any dataset. The proposed system achieved an accuracy rating of 92% for the calculated forum activity rating. The reason for the high accuracy can be attributed mainly to the method of calculation. A minor drawback of the methodology used is that if the analysis process is stopped halfway through, it cannot be resumed and must be restarted. However, this is not a major issue because the most time-consuming task is the data gathering phase.

4 CONCLUSION

Massive Open Online Courses have become a popular means of learning in a wide variety of subjects. MoocRec V1 introduced novel features to solve the problem of helping users find MOOCs. This paper shows the findings of using forum data to get insights which can in turn be used to improve MOOC recommendations. The results of the tests conducted show that the methods used in this paper are effective and feasible at helping users find MOOCs that are suitable for them.

5 REFERENCES

[1] D. Shah, “By The Numbers: MOOCs in 2018,” *Cl. Cent.*

[2] H. Khalil and M. Ebner, “MOOCs Completion Rates and Possible Methods to Improve Retention - A Literature Review Outlines - Introduction - Retention Rates for MOOCs - Why do Students Dropout of MOOCs ? - What are the techniques that increase retention in MOOCs ? - Conclusion,” pp. 1305–1313, 2013.

[3] C. Kent, E. Laslo, and S. Rafaeli, “Interactivity in online discussions and learning outcomes,” *Comput. Educ.*, vol. 97, pp. 116–128, 2016.

[4] R. Cai, J. Yang, W. Lai, Y. Wang, and L. Zhang, “iRobot: An Intelligent Crawler for Web Forums,” *Proceeding 17th Int. Conf. World Wide Web (WWW 2008)*, pp. 447–456, 2008.

[5] Q. Gao, B. Xiao, Z. Lin, X. Chen, and B. Zhou, “A high-precision forum crawler based on vertical crawling,” *Proc. 2009 IEEE Int. Conf. Netw. Infrastruct. Digit. Content, IEEE IC-NIDC2009*, pp. 362–367, 2009.

[6] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining Microblogging Microblogging = posting small blog entries Platforms,” *Analysis*, pp. 1320–1326, 2010.

[7] B. Pang and L. Lee, “Sentiment analysis using subjectivity summarization.pdf,” 2002.

[8] S. Aryal, A. S. Porawagama, H. M.G.S., and T. S.D., “MoocRec: Learning Styles-Oriented MOOCs Recommender and Search Engine Department of Software Engineering Sri Lanka Institute of Information Technology Sri Lanka MOOCREC : LEARNING STYLES-ORIENTED MOOCS RECOMMENDER AND SEARCH ENGINE October 2018,” no. October, 2018.

[9] S. B. M. Hilmy, “Coursera Course Extractor,” 2019. [Online]. Available: https://github.com/MoocRec2/Class-Central-Crawler.

[10] S. B. M. Hilmy, “Edx Crawler,” 2019. [Online]. Available: https://github.com/MoocRec2/Edx-Crawler.

[11] S. B. M. Hilmy, “Future Learn Crawler,” 2019. [Online]. Available: https://github.com/MoocRec2/FutureLearn-Crawler.

6 APPENDICES