Exploring MOOCs With Sentiment Analysis and Dropouts: What Does It Tell Us?

Mayur Mahajan

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Timothy Savage

August 2020

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

MAYUR MAHAJAN

University of Dublin, Trinity College

September 5, 2020

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

MAYUR MAHAJAN

University of Dublin, Trinity College

September 5, 2020

Acknowledgments

I would like to thank a number of people who have helped make this dissertation possible.

I am deeply thankful to my supervisor, Professor Timothy Savage, who has guided me through the formative process of this dissertation and offered me his continued support, expertise, valuable feedback and encouragement. I would also like to thank Dr. Silvia Gallagher for her guidance, recommendations and constant support for this dissertation.

I express my gratitude to thank Trinity College Dublin, the School of Computer Science and Statistics, for the recognition and opportunity to learn from the very best in the field of Computer Science.

Most of all, I am immensely grateful for my family and friends, who without their unconditional love and support I would most certainly not be here today.

Mayur Mahajan

University of Dublin, Trinity College
August 2020

Exploring MOOCs With Sentiment Analysis and Dropouts: What Does It Tell Us?

Mayur Mahajan, Master of Science in Computer Science

University of Dublin, Trinity College, 2020

Supervisor: Timothy Savage

Online education has taken a huge surge in recent years due to the flexibility and accessibility it offers in terms of learning. Massive Open Online Courses are free courses offered by various websites which are open for everyone to enrol. The structure of the course has platforms on which students can discuss, practice and learn. Even though these courses are very popular, the high attrition rate has always been a concern for the MOOC owners. In order to scrutinize these dropout rates, it is important to perceive the students' attitude towards the course and these opinions can be extracted with the help of sentiment analysis.

The main objective of this dissertation is to explore the relationship between the attrition rate of the MOOC and student sentiments obtained through the comments. In addition to the dropouts rate, the mutual association of step completion rate with the student sentiments was studied. Furthermore, the data was divided according to their step content type, and then the same analysis was repeated to study the differences between the trends in different step types.

There are several methods to choose from to implement sentiment analysis and for this thesis, the NLTK module in python together with SentiWordNet corpus was found to be the best suited for the available data. The correlation was explored with the help of pandas profiling module in python and it was found that there is no relationship between the student sentiments and dropouts rate for the data considered for this analysis. Apart from video steps, it was identified that the step completion ratio also does not have any relation with student sentiments. This thesis also tries to discover the reasons behind the outcomes. Finally, the thesis concludes with a few limitations of this process along with directions for future research.

Contents

Acknowledgments **iii**

Abstract **iv**

List of Tables **viii**

List of Figures **ix**

Chapter 1 Introduction **1**

1.1 Background and Motivation 1
1.2 Research Questions 3
1.3 System Overview 4
1.4 Thesis Structure 5

Chapter 2 Literature Review **6**

2.1 Overview of MOOCs 6
2.2 Challenges in MOOCs 7
2.3 Natural Language Processing and Text Analytics 9
2.3.1 Sentiment Analysis 9
2.3.2 Sentiment Analysis in MOOCs 9
2.3.3 Sentiment Analysis Methods 10
2.4 Sentiment Analysis and Student Attrition 12
2.5 Summary 13

Chapter 3 Methodology **14**
3.1 Design . . . 14
3.2 Extraction and Preprocessing of Data . . . 15
3.2.1 Data . . . 15
3.2.2 Data preprocessing . . . 17
3.3 Sentiment Analysis . . . 18
3.3.1 Selection of a Method for Sentiment Analysis . . . 18
3.4 Identifying the Attrition Rates . . . 19
3.5 Exploring the Relationship . . . 22
3.6 Content-Wise Analysis . . . 24
3.6.1 Content Insights . . . 24
3.7 Step Wise Sentiment Analysis . . . 27
3.8 Summary . . . 28

Chapter 4 Results **29**
4.1 Data Preprocessing . . . 29
4.2 Sentiment Analysis . . . 30
4.3 Course Attendance and Dropouts . . . 31
4.4 Correlation Outcomes . . . 32
4.5 Content-wise analysis . . . 37
4.6 Step analysis (1.5 & 2.5) . . . 41
4.7 Observations . . . 43

Chapter 5 Discussion **44**

Chapter 6 Conclusion and Future Work **46**
6.1 Conclusion . . . 46
6.1.1 Key Findings . . . 47
6.1.2 Limitations . . . 48
6.2 Future Work . . . 48

Bibliography **49**

Appendices **54**

List of Tables

4.1 Correlation Coefficients for step completion 34
4.2 Correlation Coefficients for dropouts 35
4.3 Non-linear Correlation Coefficients for Run 3 36
4.4 Correlation Coefficients for videos - Step completion 37
4.5 Correlation Coefficients for Articles - Step completion 38
4.6 Correlation Coefficients for videos - Dropouts 40
4.7 Correlation Coefficients for Articles - Dropouts 40

1 POS tags used [Liberman, 2003] 56

List of Figures

1.1 System Overview 4

3.1 System Architecture 15
3.2 Step-wise Statistics 20
3.3 Weekly Statistics 21
3.4 Video views Percentage 25
3.5 Average Sentiment Scores 26
3.6 Content-wise student attendance 27
3.7 Step-wise sentiment Score for each comment 28

4.1 Step-wise average sentiment Score 30
4.2 Total attendance for each step 31
4.3 Step-wise Dropout rate 32
4.4 Coefficient matrices for step completion rate and sentiment scores . . . 33
4.5 Coefficient matrices for dropouts and sentiment scores 34
4.6 Dropouts vs Average Sentiment Score 36
4.7 Correlation matrices for article step completion 37
4.8 Correlation matrices for video step completion 38
4.9 Video step completion rate vs average sentiment score 39
4.10 Correlation matrices for article step dropouts 39
4.11 Correlation matrices for video step dropouts 40
4.12 Word Cloud for Step 1.5 41
4.13 Word Cloud for Step 2.5 42

Chapter 1

Introduction

1.1 Background and Motivation

In the last decade, online learning has shown substantial growth all over the world as the combination of the internet and education has eased the access level for all users so that they are able to gain the skills and qualifications at their convenience. Statistics show that since 2000, the online learning industry has grown by 900% and it is expected to grow more in the coming years [Keegan, 2020]. The quality of online education has been improving over all these years and as a result, its reputation has also developed.

The term MOOC (Massive Open Online Courses) was first used in 2008 and after that, in 2012, a company named Udacity started to develop and offer MOOCs for free [Mora, 2012]. MOOCs are free, interactive and informative courses available for students all over the world and accessible over the web. There are many MOOC providers in today's world such as Coursera, edX, Udacity, FutureLearn and more. The MOOC evolution also led to online degrees and academic certifications. More than 900 universities offer MOOCs and this shows the popularity of the MOOCs [Shah, 2019]. The topics in MOOCs have a wide range from technology and business to arts and health.

Because of the interactivity and access convenience, MOOCs have a high number of students getting enrolled. This number of students decreases as the course proceeds as many students drop off from the course due to different reasons. These reasons include

but not limited to time limits, other commitments, difficulty levels, disliking course contents or only for browsing purpose. Apart from these reasons, one that stands out for all MOOCs is the self-motivation of the students and this cannot be influenced by the MOOC. These reasons either lead to an unfinished step or dropping off from the course. Despite the popularity of MOOCs all over the world, the high attrition rate remains a concern for the MOOC owners.

MOOCs offer interactivity in terms of discussion forums on which students can ask their doubts and share ideas or information with each other as well as with the course tutor. These forums are also used to have general conversations with other students about the specific course topics. Based on these posts, one can easily identify the opinions of the students at a specific step in the course. Considering the high dropout rates, it is important to know what opinions students have towards the course.

The idea of sentiment analysis is a part of Natural Language Processing (NLP) that enables the sentimental applications that are used to extract informative texts and facts. In the past few years, textual and sentimental analysis is widely used in the industry. To give an example, twitter sentiment analysis is one of the most popular methods to extract meaningful tweets. The political views conveyed via tweets can be divided into positive or negative based on keywords [Maynard and Funk, 2011]. Also, customer reviews of any product can be easily classified into different categories based on its sentiment.

As the number of comments on the MOOC is very high, it is not possible to read and understand them all. In order to extract these opinions, sentiment analysis of the forum posts can be implemented which can help MOOC owners understand the attitude of the students towards their course. Depending on these extracted opinions, MOOC owners can decide the next actions to take to help reduce the attrition rate. Although the sentiments of the students attending the course may vary, their impact on the dropouts is questionable as there can be various factors behind the dropouts rate. After calculating the sentiment of the students and the attrition rate, it is possible to explore the relationship between the two.

The dataset used for this analysis is extracted from the FutureLearn website and has statistics of a MOOC related to The Book Of Kells [FutureLearn, 2012]. The MOOC is divided into four weeks, each having multiple steps with different content types such as videos, articles, quizzes, exercise or discussion.

1.2 Research Questions

This section outlines the research questions that this project is engineered to tackle and also describes the possible hypotheses. Also, system flow overview is given at the end in order to give an idea on how the research questions are investigated.

Question 1 Does the student sentiments based on the comments have any association with the dropouts rate or step completion rate in a MOOC?

• Before running the analysis on the data, it is important to prepare it for the analysis as it cannot be processed directly.

• It is important to find out which sentiment analysis approach is best suitable for the available data as there are various methods to go for.

Description

Considering the high dropouts rate, it is important for the MOOC owners to understand the trend and restrict the falling numbers in attendance. The opinions of the students can only be extracted with the help of the comments they provide but to check if these comments are useful in identifying this attrition rate, their relationship with dropouts has to be explored.

In this process, two of the main tasks are data preprocessing and sentiment analysis. The collected data has to be cleaned before the sentiment analysis.

Question 2 Does the dropouts rate vary based on any particular content type of the step of course?

• Do sentiments of the students change drastically with respect to the content type of the step?

• Is there any particular step having a very low or very high sentimental score that differentiates it from other steps? If yes, then why?

Description

It is important to know if the students do not prefer any particular content type which results in higher dropouts than other steps. With the aim of reducing dropouts, its relation with sentiments of students for these different content types is recognised separately. To check if any particular step is most popular or least famous amongst students, sentiment scores of all the steps are considered and the variance of sentiments is checked. If any of the steps has a score that sets that step apart, the reason behind that score is analysed.

Below are the key hypotheses which the dissertation investigate:

• Hypothesis 1: The average sentiment score for a step is correlated to the attrition rate in a MOOC.

• Hypothesis 2: The average sentiment score of a step affects the step completion rate and they are associated with each other.

• Hypothesis 3: The correlation between student sentiments and attrition rate has a significant difference for different content types.

• Hypothesis 4: The correlation value between step completion rate and sentiment scores vary for different content types.

1.3 System Overview

The system consists of three main modules, data preparing, sentiment analysis and exploring the association between sentiments and dropouts. After receiving the data from the website, the NLTK module of python is used to preprocess the comments before passing them as an input to sentiment analysis. Sentiment analysis is implemented with SentiWordNet as a lexicon in Python NLTK. Before carrying out the sentiment analysis, two more methods were taken into the account. The dropouts and sentiments are given to the pandas profiling as input so that the correlation between the two is calculated. Figure 1.1 gives an overview of the system.

Figure 1.1: System Overview

Apart from these three, there are a few processes such as content-wise analysis and interpretation of the results that will be implemented in this project.

1.4 Thesis Structure

After outlining the primary goals of the dissertation and summarizing the system processes, this section gives an overview of how the thesis work is organized into different chapters:

Chapter 2: The next chapter, i.e, chapter 2 is focused on the literature review around the MOOCs, its characteristics and sentiment analysis methods. It also discusses the problems around MOOCs and previous research into this area. As this project is majorly based on sentiment analysis and student attendance in MOOCs, in-depth literature about the two is presented in this section.

Chapter 3: This section describes the system architecture in detail and the design of all technical processes is described. The tasks implemented as part of data preprocessing are elaborated in detail. The process of selecting the method for sentiment analysis is discussed in this section. Also, the calculation of dropout and correlation coefficients are explained with the help of formulas.

Chapter 4: The evaluated results are represented in this section. With the help of the correlation matrices and the tables describing the coefficient values, the outcomes of the analysis are explained. Also, the results are interpreted efficiently to answer the questions of hypotheses.

Chapter 5: This chapter discusses the possible explanations of the answers to the research questions and the most significant outcomes obtained from the analysis. Interpretation of the results from this analysis is compared with the previous research.

Chapter 6: This final chapter concludes the dissertation and also gives recommended directions for future research and improvements.

Chapter 2

Literature Review

This section of the dissertation is intended to provide a background to the problem description and introduces different concepts based on the previous research. The prior research into the area of MOOCs, challenges in MOOCs, and sentiment analysis is covered in this chapter. And finally, a strong motivation for this dissertation is mentioned.

2.1 Overview of MOOCs

In recent years, online learning has become popular in higher education than ever due to the rise of digitalization and the interactivity they offer [Min et al., 2019]. To give detailed information about any of the curricular domains or any discipline, many educational organizations offer free online courses for learners. 'Massive Online Open Courses (MOOCs) represent free, self-paced, open-access, global, instructional and well-structured courses [Baturay, 2015]. Types of steps in a MOOC can be videos, articles, problem sets and forums. MOOCs also offer discussion platforms for interaction either between students taking the course or between students and the course organizers. These discussion platforms can be forums on the same site or some external websites. These are often associated with other content types such as Videos or Articles but these are completely different steps which are not informative.The comments section on videos or articles are not discussion platforms offered by MOOCs. Some of the MOOCs also have some problem sets such as multiple-choice questions or

assignments at the end of sections.

According to Shah [2019], MOOCs have around 110 million learners across more than 900 universities and in excess of 2500 have been launched in 2019 alone. Also, based on the offerings and number of users the top MOOC providers are Coursera, edX, Udacity, FutureLearn and Swayam [Shah, 2019]. FutureLearn enables its learners to discuss their doubts or learnings with other students on the go in order to make the learning an enjoyable experience [SocialLearning, 2015]. FutureLearn offered the most number of online degrees with MOOC in 2019 amongst all the MOOC providers [Shah, 2019].

2.2 Challenges in MOOCs

Although MOOCs do offer a lot of flexibility in terms of learning, there are several issues that students or course owners face. As part of their research in this area, Fournier and Kop [2015] categorize the areas related to the MOOC challenges and and explain them in detail. According to this research, the learning experience, personalization, individual learning needs, and ethics are some of the broad fields that involve MOOC issues [Fournier and Kop, 2015]. K. F. Hew & Cheung [2014] in their experiment divide the challenges for learners and teachers based on their experiment. Challenges for students include lack of incentive, lack of focus on forums, insufficient prior knowledge, or shortage of time and on the other hand, tutors face issues such as lack of response in online discussions, monotonous teaching, or demands of money [Hew and Cheung, 2014].

Considering these challenges, the major concern regarding MOOC is the low completion rates of the course [Sharkey and Sanders, 2014][Chen and Zhang, 2017]. The loss of students in MOOC affects the development in MOOCs in a negative way and also impacts the MOOC standards [Yin, 2019]. In addition to this, Y. Chen & Zhang [2017] implies that future developments in MOOC can get restricted due to attrition. On the other hand, a few authors do not consider high attrition rate as an obstacle or a problem for MOOCs [Balch, 2013] [Iraj, 2014] [Reich, 2014]. Balch [2013] considers the attrition rates as a bogus argument because of the investment or commitment level made by the students differ from the regular courses. Iraj [2014] also does not find the low completion rates problematic as they do not consider many use-cases of drop-outs

such as trial and error. Reich [2014] believes that dropout rates are mainly dependent upon student intentions.

Many students drop-out at some stage of the course due to various reasons and hence the attendance of the course decreases gradually [Chaplot et al., 2015]. These reasons may include lack of motivation, course content, language barriers or failure of gaining expected practical knowledge. Only a few MOOCs provide certifications that can be used for academic or professional purposes and in such cases, learners do have the motivation to complete the course. Individual attrition probabilities can be considered with the help of a neural network model to personalize intervention in a MOOC [Xing, 2019]. Observing student's performances during the course and provide suggestions to them can be useful to retain students throughout the course [Widyahastuti and Tjhin, 2018].

Sunar et al. [2017] in their findings on learners engagement in MOOCs focus on the activities of students such as posts or following someone. The results suggest that the students who tend to follow someone have a higher probability of completing the course [Sunar et al., 2017]. The completion rate also depends upon the start date, length of the course and its type of assessment [Jordan, 2015]. Based on their findings, Coffrin et al. [2014] with the help of visual analytics suggests that there should be some informal diagnostic test before starting the course so that students are clear about the idea of MOOC. This would solve the issues such as lack of motivation or disliking content and as a result, the attrition rate would go down [Coffrin et al., 2014]. A study done by Deshpande & Chukhlomin [2017] indicates that visual design or self-assessment do not impact the completion rates but accessibility, interactivity and the contents of the course do have a significant influence on student attrition. These factors also have a positive impact on students motivation towards learning [Deshpande and Chukhlomin, 2017]. Ferschke et al. [2015] based on their experiment using survival models suggests that the collaborative chat slows the attrition rate. This collaborative work can either be a reflection exercise or a chat [Ferschke et al., 2015]. The forum posts play an important role in motivating student during the course and also contributes to the completion of assignments on time [Zhang et al., 2018].

2.3 Natural Language Processing and Text Analytics

Considering the advances in internet technology, data is generated at a faster pace than ever before and the quantity of this data is huge. The data generated in articles, blogs, newspapers and social media is unstructured and part of textual data. This textual data has to be preprocessed before any analysis. This assessment and classification of the data come under text analytics and it is known as Natural Language Processing. One of the most popular tasks in Natural Language Processing is Sentiment Analysis.

2.3.1 Sentiment Analysis

Based on the previous research, one of the ways to gain information about the dropouts can be by analyzing the opinions of the learners about the course based on the discussion forums. To analyse student satisfaction based on the contents of a MOOC, the sentiment analysis of the discussion comments can be useful [Wen et al., 2014]. To make any text meaningful or to extract any useful information out of it, the text needs to be processed. When this textual analysis is performed to gain a better understanding of the attitude of the author of the text or to find out their opinion through the text, it is known as the Sentiment analysis. Applying natural language processing to mine some hidden value from a variety of information through text is part of sentiment analysis. One can define the attitude or emotion of the author with the use of sentiment analysis in terms of positive, negative or neutral. Sentiment analysis is widely used in different disciplines nowadays due to its business and society values.

2.3.2 Sentiment Analysis in MOOCs

Interpretation of the student opinions can be useful for various purposes such as student participation, performance, dropouts or for deciding how effective the MOOC is. Student interactions are classified by Harris et al. [2014] with the help of sentiment analysis to check their involvement in the course. This will help MOOC authors to identify the changes required and also help in tracking issues [Harris et al., 2014]. Another work in this area is done by Kumar & Jain [2015], that collects feedback from

the students and uses that for sentiment analysis. The feedback is divided into subjective and objective sentences and then opinions are extracted from subjective ones [Kumar and Jain, 2015]. One of the ways to define MOOC success can be the combination of course features and perception of students that can be obtained by analysing their opinions [Hew et al., 2020]. Teaching patterns or strategies also can be improved based on the attitude of the learners [Liu et al., 2016].

Another aspect in this area includes the understanding of social interaction of students based on their skills, sentiments and social quotient [Moreno-Marcos et al., 2018a]. One of the easiest techniques can be to classify the student opinions into favourable or unfavourable towards the course and calculate the ratio [Suzuki et al., 2006]. A different approach in this research area is given in Peng & Xu [2020], the emotions of two groups as completers and non-completers are compared and their behaviour is analysed statistically. This analysis gives the overview of differences between both groups so that non-completers can be analyzed [Peng and Xu, 2020]. Wang, Hu,& Zhou [2018] in their work on improving MOOC teaching proposes a semantic analytics model that first divides the learners based on their participation and then evaluates the emotional qualification to promote personalized teaching in MOOC. This model also predicts the probability of graduation for students so that it can be improved [Wang et al., 2018].

2.3.3 Sentiment Analysis Methods

There is a wide range of methods to extract sentiments from any text and most of them use the English language for processing. Implementation of these methods can be classified into supervised and unsupervised learning-based methods.

Supervised Methods

The supervised approach is a type of machine learning technique that uses a training set and a test set that needs to be classified. Most popular techniques for classification in supervised learning for sentiment analysis are Naïve Bayes and Support Vector Machines.

To classify the user's sentiments into six different categories, used twitter API to extract the sentiments of the students and then users were given a chance to classify the incorrectly placed sentiments for better results [Harris et al., 2014]. Naïve Bayes

classifier is applied to the data by Kumar & Jain [2015] to classify the sentiments into different categories. A new approach of gradient boosting is used by Hew et al. [2020] to classify opinions into different classes. These factors are divided as learner and class levels and describe satisfaction in MOOCs [Hew et al., 2020]. By simply subtracting scores of negative words from scores of positive words, sentiment score was calculated that indicates the sentiment level of the sentiment [Shapiro et al., 2017]. After dividing the comments into different behavioural categories, Gibbs sampling was used by Peng & Xu [2020] in their investigation on behaviours of learners to get emotions and then they were compared between sentiments of completers and non-completers of the course. Liu et al. [2016] make use of SVM and logistic classifier in their research to build the model for sentiment analysis after selecting features using particle swarm optimization model.

Unsupervised Methods

This method of machine learning uses predefined lexicon to classify the sentiments of the users. It is considered to be the most transparent method of sentiment analysis. The unsupervised approach can be carried out using human annotators or dictionary-based method. Human annotator's method includes creating a lexicon by the linguistic experts by manual labelling the words and on the other hand, the dictionary-based method uses a prewritten dictionary that has a word bag with defined sentiments which can be used for classification.

SentiWordNet is an open-source lexicon resource that can be used for sentiment analysis [Baccianella et al., 2010]. Estrada et al. [2020] created two corpus SentiText and eduSERE to classify the emotional opinions and got an accuracy of 93% and 84% respectively. Along with positive and negative, a few different classes such as engaged or excite were added to these corpora [Estrada et al., 2020]. In addition to this, WordNet, TextBlob, Stanford NLTK, and VADER have their corpus that can be used for the sentiment analysis. Moreno-Marcos et al. [2018] calculates sentiment scores using SentiWordNet, the random forest is applied for the classification. These models are used to extract a bit complicated emotions such as frustration or excitement [Moreno-Marcos et al., 2018b].

Benefits of Sentiment Analysis and its Relevance in MOOCs

Below are some top benefits of sentiment analysis [Marta, 2019]:
• Customer experience: Understanding the audience and aim the message accordingly. In MOOCs, it is important for MOOC coordinators to understand the feedback of the learners and work on it for improvement.
• Marketing strategy: Understanding the motivations of the customers and the intent of their search. This can help in planning a unique marketing strategy. To reach out to many students, it is important to know the intent of students who take up the course initially.
• Brand perception: Based on the reviews or comments, it is possible to know the attitude of the audience. In order to check the popularity of the course amongst students, it is good to know the views of the students.
• Boosting service: Monitoring negative keywords or phrases can help solve critical situations. By detecting the relevant negative words from comments, the root cause can be identified for them and it can be resolved to upgrade the MOOC.

2.4 Sentiment Analysis and Student Attrition

As stated earlier, motivation and understanding of the content affect the attrition rates and it is easy to perceive these factors with the help of sentiment analysis of the forum posts. Adamopoulos [2013] in their research on student retention in MOOC conclude that student sentiments do have a positive impact on the successful completion of the course. Based on the results, insights on further actions are also elaborated [Adamopoulos, 2013]. Any type of participation in discussion forums can restrict students from dropping out of the course and any interaction may result in retention [Chen and Zhang, 2017]. Frequent and repeated interactions on the discussion forums result in lower dropout rates in MOOC [Sunar et al., 2017]. Alblawi & Alhamed [2017] predicts student performance in online learning based on the sentiments of the students along with their personality attributes and gets accurate results [Alblawi and Alhamed, 2017]. One of the important feature to predict dropouts in MOOCs is Sentiment Analysis of the comments on forum [Chaplot et al., 2015]. Kagklis et al. [2015] attempts to evaluate students success with the course based on their

attitude towards the course based on the sentiment for the duration of the course. The overall education process can be enhanced by the tutors if they consider the interaction pattern and student sentiment [Kagklis et al., 2015].

2.5 Summary

To summarize, MOOCs are the free open-access courses offered by educational institutions and the biggest challenge for them is low completion rates. To tackle this problem, there have many prediction methods that have been implemented so that it is possible to retain students. Sentiment analysis of the discussion forum posts is used to obtain the opinions of students so that course owners are aware of the unhappy students.

There has been considerable research on predicting dropouts in MOOCs with the help of different features [Chaplot et al., 2015] [Chen and Zhang, 2017]
[Sharkey and Sanders, 2014] [Xing, 2019]. And also on the sentiment analysis of the discussion forum posts in MOOC [Cobos et al., 2019] [Estrada et al., 2020]
[Ferschke et al., 2015] [Kumar and Jain, 2015] [Moreno-Marcos et al., 2018a]. To get more insights into the two, exploring the relationship between sentiment analysis and student attrition in MOOCs needs more research and testing. It is important to conduct this study to check the impact of discussion forum posts on the dropouts or step completions.

Chapter 3

Methodology

This chapter is intended to give an overview of the processes in the system implementation. As stated earlier, the system mainly consists of three stages and all of them are elaborated in this chapter. Data preprocessing is required to deal with unstructured data and to prepare data to mine the opinions of the authors of the comments. Sentiment analysis extracts the opinion of the author and then it can be compared with the dropouts and step completion rate to find the relationship between the two.

3.1 Design

The design of this implementation includes data preprocessing tasks such as Tokenization, POS Tagging, Lemmatization & stemming, and Stop words removal, sentiment analysis of the forum posts using SentiWordNet and finding a correlation between the student dropouts and their opinions using the pandas profiling in python. The diagram below explains the detailed architecture of the system. The primary task of this thesis starts with collecting the data for the Book of Kells: Exploring an Irish Medieval Masterpiece course, a MOOC offered by Trinity College Dublin. As part of textual analysis, the data needs to be preprocessed so that the learners' opinions can be calculated easily with the help of a lexical resource. After calculating the sentiment values of the learners, the correlation with the weekly dropouts is calculated.

Figure 3.1 shows the architecture of the proposed system with detailed components that are essential in the process of exploring the data.

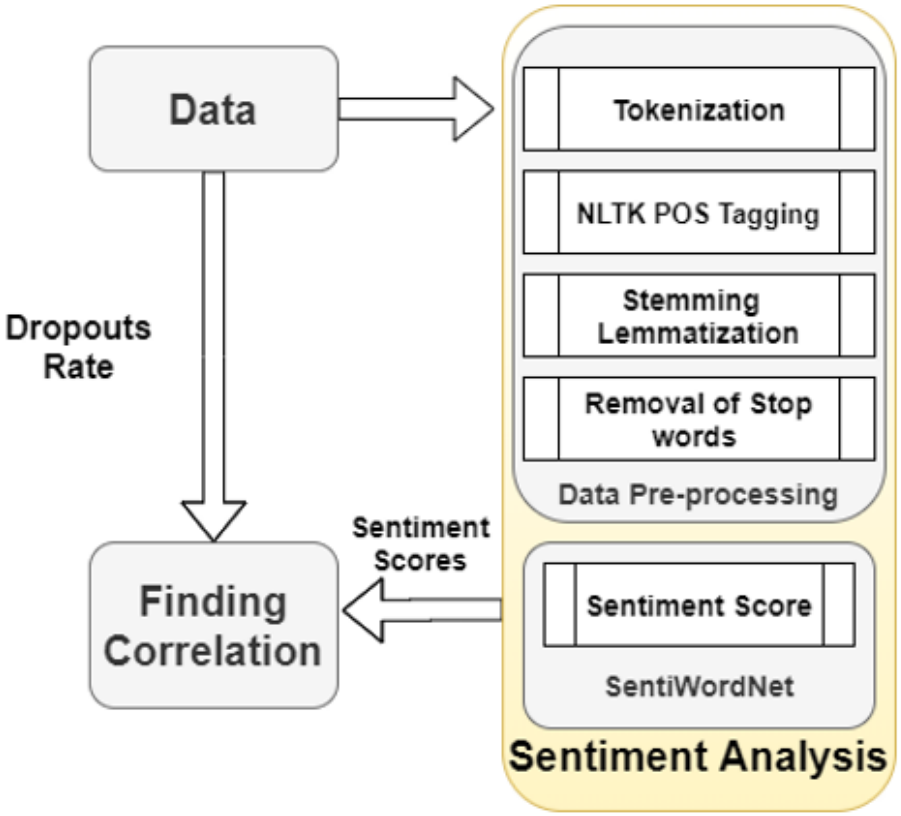

Figure 3.1: System Architecture

3.2 Extraction and Preprocessing of Data

The details of the tasks involved in the data extraction and cleaning are covered in this chapter.

3.2.1 Data

The data is collected from FutureLearn [2012] website and consists of all the statistical information about a MOOC named Book of Kells: Exploring an Irish Medieval Masterpiece. There are three different runs of the MOOC and for each run, the data is separately extracted from the website. The MOOC steps are divided into four weeks. The content types in the MOOC are Videos, Articles, Quizzes, Discussions and Exercises.

The data for every run is divided into three tables namely, videos, activities and comments. The videos table contains the statistics about every video step of the course, the activities table has information about the completion dates of learners whereas the comments table has all the forum posts with respective learner IDs. Data from all three tables are important for the analysis at different stages. As the data is available for three runs for the same MOOC, the analysis has to be repeated for every run. For sentiment analysis, the comments table is used while to check and compare the dropout rates, activities and videos tables are used.

a) Videos Data – Out of the 55 steps of the MOOC, there are 14 videos distributed over the period of four weeks. This table of Videos consists of all the viewing statistics of these 14 steps such as the number of views, total downloads, viewing percentages, device information and regional views. This table is helpful in analyzing the viewers inclination towards a particular set of videos and their preferences while watching the videos.

b) Activity Data – This table contains the information about the start and completion date for every user that logs in to that particular step. The data present in this table gives activity details of all users for every step. Based on this table, for every step, one can easily identify the number of users who have completed that particular step or left it incomplete. The total number of students who joined the course was close to 6000 for each run.

c) Comments Data – This table has all the comments on discussion forum for every step and also has the information on comment's author ID, timestamp, likes, replies, and alterations to comments if any. On average, there are around 20,000 comments for each run of the MOOC. Learners' sentiments towards that particular step can easily be extracted with the help of these comments and these comments also depicts the users level of interaction.

To understand this MOOC dataset along with its attrition rate and author sentiments, exploratory data analysis is very important. Exploring data can be considered as one of the primary tasks to acquire insights from data [Yu, 2010]. Below are the steps to be performed to achieve the goal of exploring the dataset.

3.2.2 Data preprocessing

In order to extract the opinions of learners, the sentiment analysis on the comments posted on the discussion forum for every step is performed. To prepare the data for the analysis, some normalization tasks have to be performed on it. Before proceeding to the normalization tasks, the data needs to be divided into tokens and cleaned properly.

Tokenization

The process of splitting the complete sentences into tokens is known as tokenization. This divides the strings based on the whites spaces and punctuation. These tokens may contain noise in the form of punctuation, hyperlinks or special characters. This noise has to be removed so that only meaningful text is processed. The noise is removed with the help of "string" library in python.

As these words are can be any form as -ed, -ing or -s, they have to be converted to their root form by normalizing them. To extract the sentiments out of these words, they must be present in their basic forms. Normalization also helps in grouping the words that have the same meaning. Following are the steps involved in normalization:

Part Of Speech Tagging

After the strings are divided into tokens, it is easy to tag these tokens based on their parts of speech. This process is helpful to understand the context of every word from the input data. This tagging is based on the relative position of the words in the respective sentences and this algorithm is called as the "POS Tagging". Generally, the word tagged with NN is a noun and it is a verb if it starts with VB.

Stemming and Lemmatization

It is difficult to process the words with affixes and therefore stemming, a process that removes the affixes from the words is used. Stemming transforms the words to their root form. Lemmatization is used to analyze the basic structure and parts of the words and then return their dictionary form, known as a lemma. Even though this process was very useful and helped the accuracy, it made the overall process slow.

Removal of Stop Words

Before processing the data, the last step is to filter out the stop words. All English words are listed with the help of the words present in a corpora named 'stopwords' in NLTK and then the words that are not present in this corpora are discarded. Also, words having a length of less than three are discarded as they do not give any particular meaning to this data.

3.3 Sentiment Analysis

Following normalization, the next stage of the analysis is to calculate the sentiment score for the words in comments. As this analysis is based on the data of the course on the Book of Kells, there are no preexisting patterns or labels of the data which can be useful for the supervised learning methods. As a result, due to the lack of training data, unsupervised learning method was the only choice in this analysis. As there are a few steps of the MOOC that are quizzes, these steps do not have any comments on the forum. This resulted in discarding the four steps, i.e, 1.13, 2.14, 3.14 and 4.13, one each from four weeks as there were no comments on the forums of these steps.

3.3.1 Selection of a Method for Sentiment Analysis

Based on the previous research on sentiment analysis, the most popular choices for the unsupervised learning in python were Standford NLTK, TextBlob using Naïve Bayes Classifier and NLTK using SentiWordNet [Estrada et al., 2020][Esuli and Sebastiani, 2006] [Moreno-Marcos et al., 2018b]. Firstly, to find out which method is efficient and best suitable for the available data, a chunk of data, i.e, few of the comments were selected randomly and these methods were used on them.

Alternative Techniques

As the purpose of this analysis is to find out the correlation between the sentiments and dropouts of the course, the results of sentiment analysis from the implementation using Stanford NLTK were not of much relevance. It divides the sentiments into different categories such as Very Positive, Positive, Neutral, Negative and Very Negative rather

than giving the exact scores for each comment. The other two methods, TextBlob and NLTK using SentiWordNet give a numeric score for every comment which seem satisfying for this analysis. TextBlob gives the classification category with the help of numeric polarity and subjectivity scores of the text but the overall sentiment score of the comment is not specified.

SentiWordNet

The implementation of NLTK module of Python using SentiWordNet corpora overcame these limitations and returned an overall numeric score for each comment. Also, all the processes are transparent in this method and it preprocesses the data quite efficiently. It is the most straightforward method for opinion mining and consequently, it is highly popular [Chen and Sokolova, 2018].

SentiWordNet takes a single word and its part-of-speech tag as an input and then assigns a positive and negative score to each synset from the dataset. Synset is an interface from NLTK that is used to check the words in the WordNet. If the word is not present in the synset, the score for that particular word is returned as null. To get the overall sentiment for each comment, the negative score for each comment is subtracted from the positive score.

Sentiment Score=Positive Score-Negative Score

Based on the overall score, extraction of the opinion for the student who has commented can be easily depicted. And therefore, as NLTK library with SentiWordNet gives an overall sentiment score that can easily be correlated with the dropouts or step completion rate, it was selected as a Sentiment Analysis method for this analysis.

3.4 Identifying the Attrition Rates

The next phase in this analysis is to gain insights on the course completion rate and percentage of dropouts so that the association with the forum posts can be explored. As the course is divided over four weeks, each having multiple steps, the attrition rate needs to be calculated separately for every step based on the total attendance of previous step and attendance for current step.

Below plots explain the MOOC data pattern in terms of student attendance for every MOOC step. The bar plot in figure 3.2 shows the number of students from run 3 of the MOOC that complete the full step and leave the step in between. This data from the article table can be useful to identify which of the stages of the course are most popular amongst the users based on the stepwise completion rates.

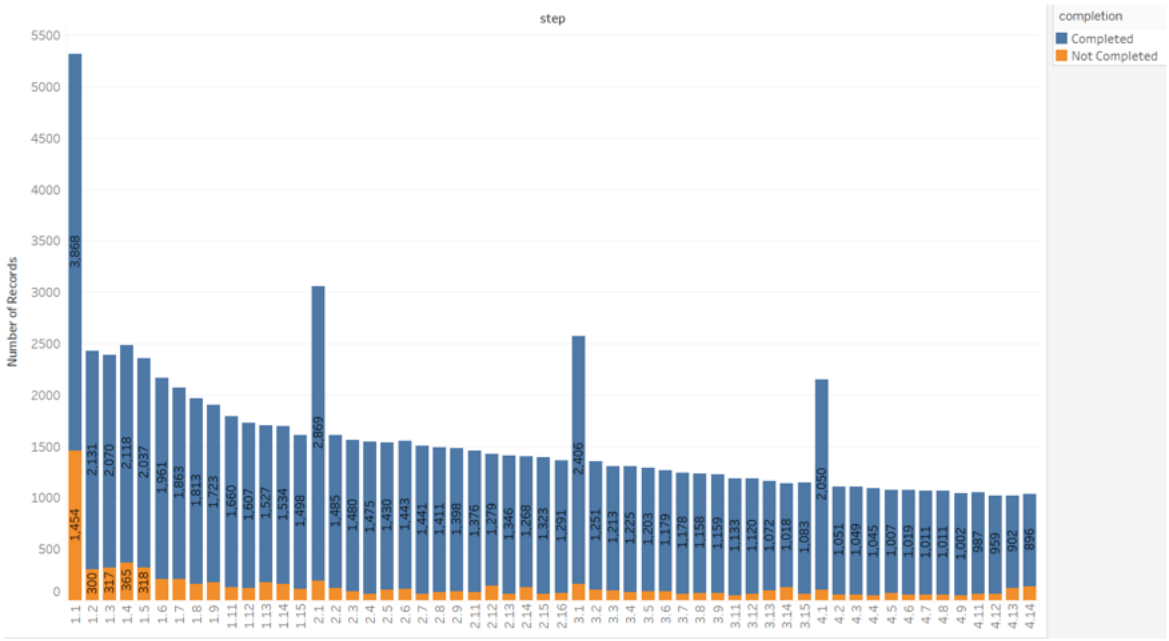

Figure 3.2: Step-wise Statistics

Coming to the weekly analysis, figure 3.3 shows the line graph of overall weekly statistics for run 3 of the data, where it can be seen that as the course progresses, the number of students who drop out from the course also increases. The weekly statistics of videos show the attrition in the number of learners over a period of four weeks and it is perceivable that there is a significant drop in attendance of the course. The plot shows that more than 47% of the total students who started the course dropout as the course goes forward.

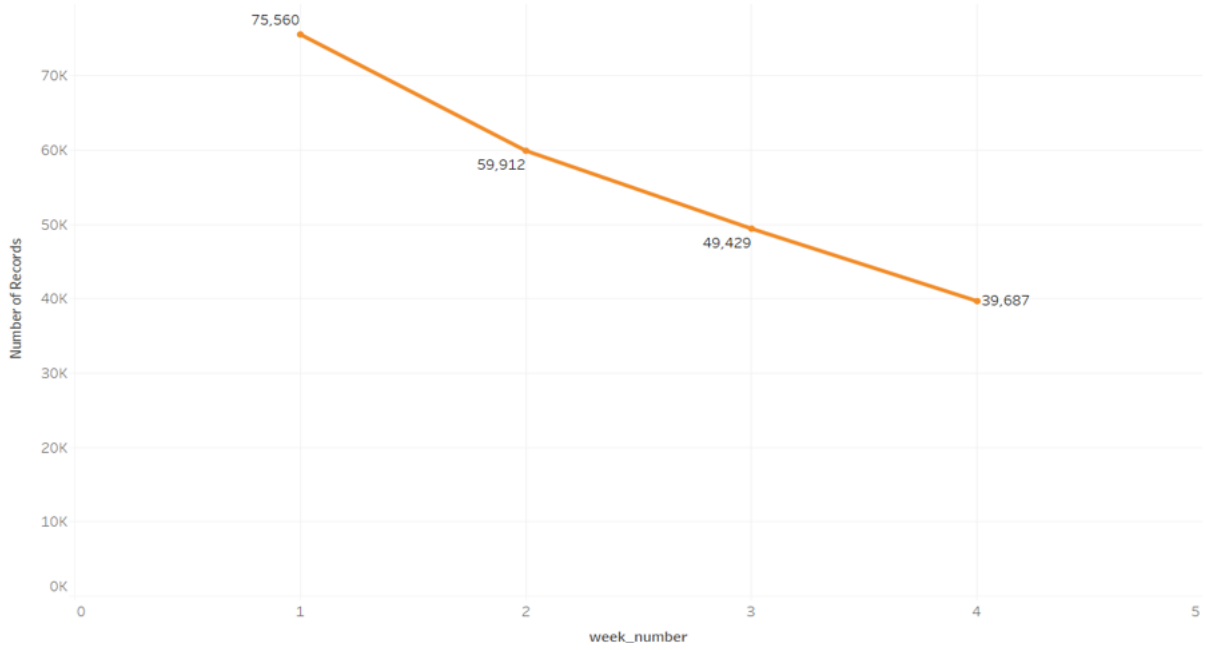

Figure 3.3: Weekly Statistics

The completion rate for every step is the percentage of total viewers who have finished the step and it is calculated as –

$$Completion\_percentage = \frac{Numberofstudentsthatcompletethestep}{Totalnumberofstudentthatstartthestep}$$

Dropouts for a course are the students that leave the course while it is in progress and never complete it. It is calculated as-

$$Dropout\_rate = \frac{Totalstudentsforpreviousstep - Totalstudentsforthisstep}{Totalstudentsforthepreviousstep} * (100)$$

After the calculation of completion percentage and dropout rate, the data for the first step of each week was discarded as the number of students attending these steps were much higher than the other steps. As we can see from figure 3.2, these steps have recorded larger audience than others and this can mislead the analysis. The reason behind this change in attendance is many students tend to view the contents of the

week by just going through the first video and then deciding if they want to proceed or not based on the initial video. At this stage, all the data required for the analysis is complete and can be proceeded to the main objective of this dissertation, i.e., to explore the association between attrition and sentiment scores.

3.5 Exploring the Relationship

The next phase of this analysis is to explore the mutual relationship between the step completion percentage and the average sentiment of learners for the discussion forum posts for that particular step. Firstly, after getting the sentiments for all the comments of the posts, the average sentiment is calculated for that step by summing up the scores for each comment and diving them by the total number of comments.

$$Sentiment\_score_(avg\_per\_step) = \frac{\sum SentimentScoresofallcomments}{Totalnumberofcomments}$$

The average sentiment scores are calculated for every step for all four weeks. Before the correlation is calculated, the sentiment score, dropout, and percentage are scaled from 0 to 1 with the use of Min-Max scaling in python. The scaling is done with the help of the formula below[Raschka, 2014]:

$$Var_(scaled) = \frac{Var - min(Var)}{max(Var) - min(Var)}$$

To investigate the mutual connection between the two, pandas profiling is used. Pandas profiling is a python module that enables to perform exploratory analysis and generates HTML reports based on the input data [Amrrs, 2018] [Brugman, 2020]. The report includes visualizations such as boxplots and histograms along with all the statistical data. These statistics include unique & missing values, quantile statistics, descriptive statistics, most frequent values and correlations between variables.

The average sentiment and completion percentage for that particular step is passed to the pandas profiling module and the first report is generated. The correlation values

that are available in the report are Pearson's r, Spearman's ρ, Kendall's τ, and Phik ϕk. Pearson's coefficient measures how strongly the two variables are associated with each other whereas the spearman's coefficient uses the monotonic function to give a measure of the relationship of two variables even if data is not normally distributed. Similar to the Spearman's coefficient, Kendall's rank coefficient gives a measure of ordinal association of two variables. Phik, on the other hand, is built on top of Pearson so that it also measures the non-linearity of the two variables.

These coefficients are considered for this analysis as they give the strength of the relationship two independent continuous variables. The linear or non-linear trend in the given variables can be identified with the help of these coefficients. To give an example, for two variables x and y, Pearson's correlation is calculated as [Singh, 2019]:

$$Pearson's(r) = \frac{N \sum xy - (\sum x)(\sum y)}{max(Var) - min(Var)}$$

The correlation values for Pearson, Spearman, and Kendall are between -1 to +1 where value close to +1 indicates that the variables do have a positive correlation and values near to -1 indicates that there is a negative correlation between the two variables. For Phik, the value is in the range of 0 to 1 where 1 signifies positive association. If the correlation value for any of these correlation coefficients is close to 0, there is no correlation between the two variables. As pandas profiling gives a detailed report and calculates all four coefficients easily, it is the preferred over all other correlation techniques. Pandas-profiling also enables to validate the relationship between two variables by validating all the possible relationships, i.e, linear, non-linear, monotonic, and ordinal. An HTML report is obtained after running the pandas profiling on the two variables.

The mutual association between the sentiments of the students and dropouts is explored in the same way as with completion percentage. For this analysis, the calculated dropout rate is passed as an input to the pandas profiling along with the average sentiment score of the previous step. The second HTML report generated by the pandas profiling based on dropouts also has all the correlation matrices for all the coefficients.

3.6 Content-Wise Analysis

After running the analysis on full data, to study if the trends change based on the content type of the step, the relationship of step percentage and dropouts with sentiment scores is explored for these content types separately.

On a broader level, this Book of Kells MOOC has five content types, i.e, Videos, Articles, Quizzes, Exercises and Discussions. The informative steps of the MOOC are divided into two distinct categories based on the type of medium they use to deliver this information to students. These two steps are articles and videos and hence student dropout rate for these steps is important than the quizzes, exercises and discussions. The other three content types in MOOC, quizzes, exercises and discussions, are focused on interactions, assignments and conversations for students. And as these content types are for their practice rather than learning, they are not included in the analysis. Firstly, content-wise insights are gained to see the difference between the two.

3.6.1 Content Insights

The data available from the videos table can be used to get insights into the video views and information about viewers that can help analyze the low correlation values. Data for MOOC run 3 is considered for these insights.

According to the figure 3.2, the viewing percentage for each step gradually decreases as the course moves ahead. To focus on this, video viewing statistics are divided by the view percentage of learners. The graph below highlights the exact stage where the students' drop off from the video.

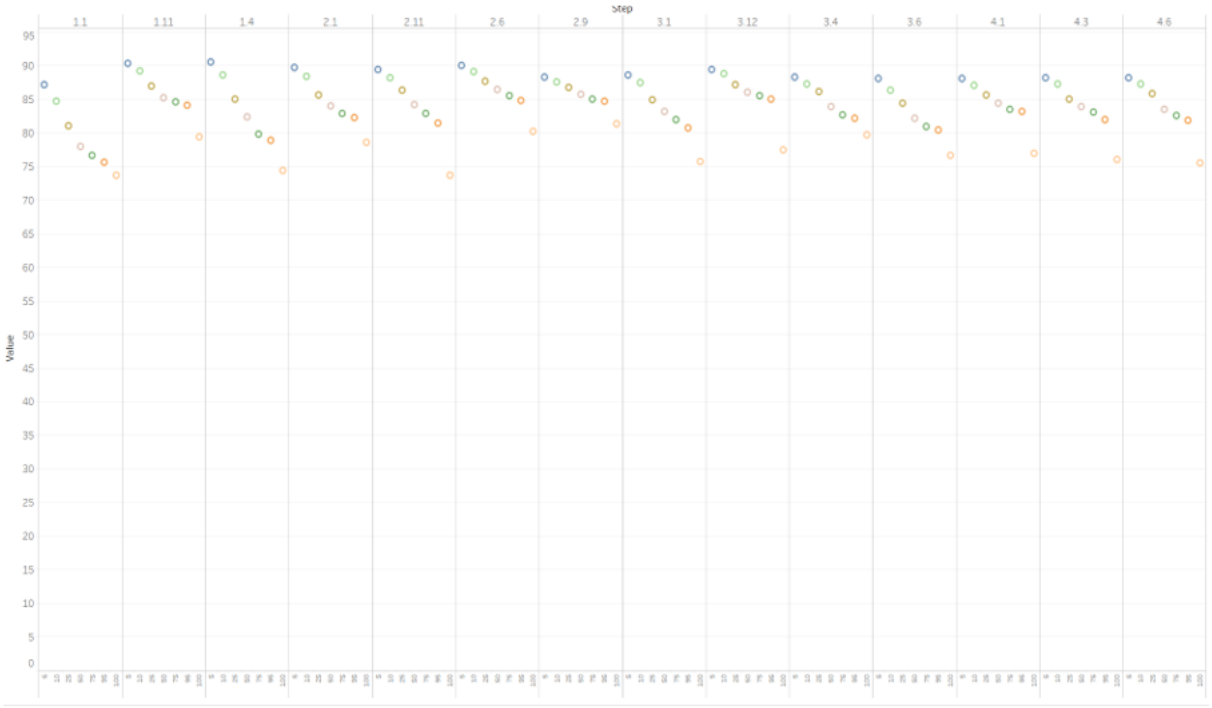

Figure 3.4: Video views Percentage

To compare the average sentiment scores in articles and videos, they are plotted on a graph and their range is analyzed. It can be observed in figure 3.5 that most of the video steps have average sentiment scores less than or equal to the neutral score whereas the majority of the article steps have better average sentiment scores than the neutral range.

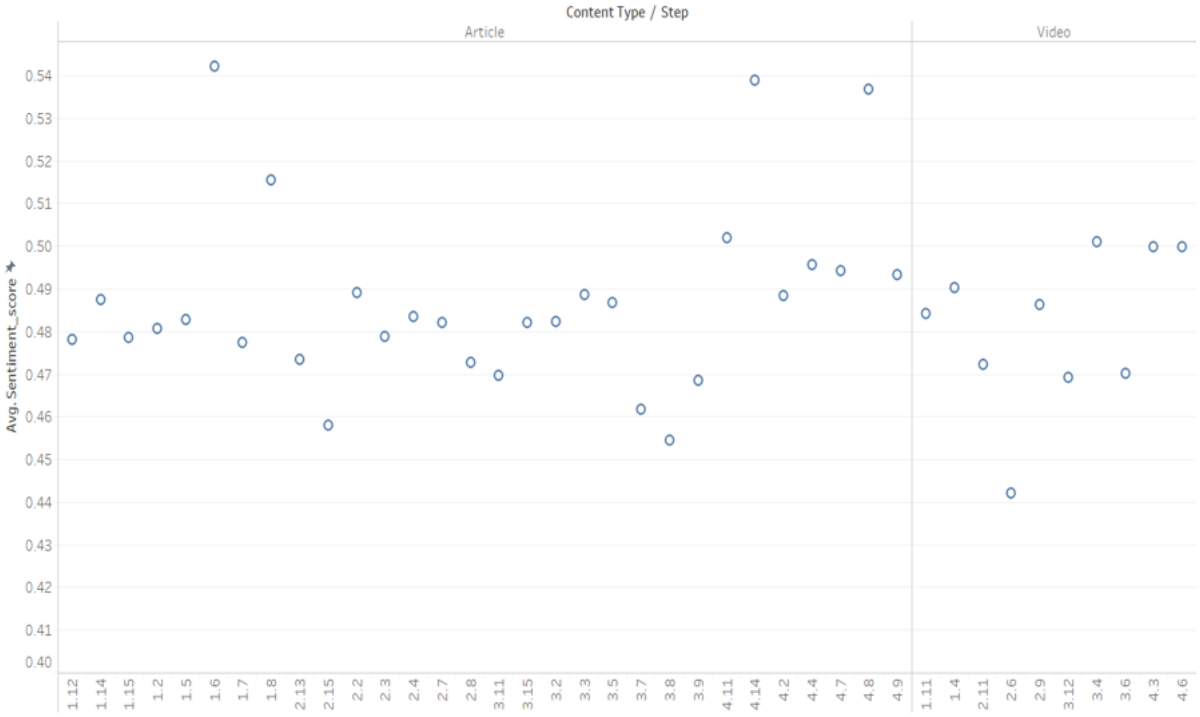

Figure 3.5: Average Sentiment Scores

The line plot below gives an idea of the total number of students attending the article and video step. The attendance for both the article and video steps goes down as the course progresses. It is clearly visible that there is a drastic drop in the attendance for the videos whereas the student population in article steps seems going down gradually.

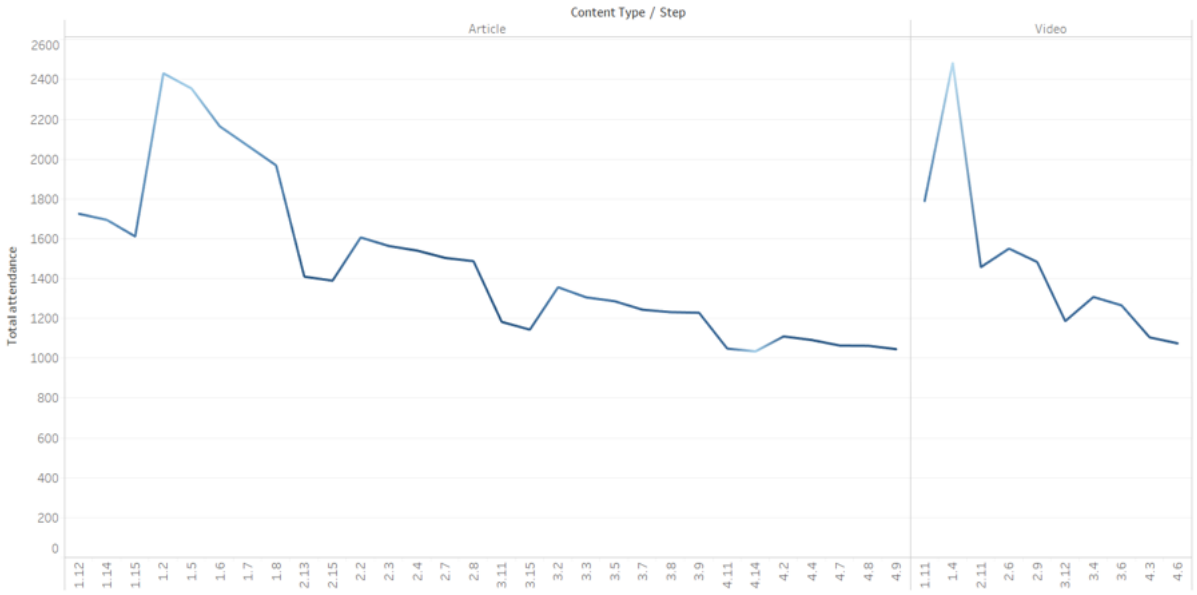

Figure 3.6: Content-wise student attendance

The next part of this analysis is to explore the correlation of sentiment scores and dropouts based on their type of content. Similar to the first analysis, pandas-profiling is used to carry out this analysis and calculate the coefficient.

3.7 Step Wise Sentiment Analysis

Total sentiment score for all comments of the step is calculated by summing them up. The distribution of the step-wise sentiment scores for MOOC run 3 is explained in figure 3.7. To check if any of the steps from the course is showing different trend in all three runs of the MOOC, the sentiment scores for every comment on the steps is compared.

As the plot below suggests, steps 1.5 and 2.5 have many comments having a score varying from the average. The scores of step 1.5 comments are more inclined towards positive sentiments whereas many of the step 2.5 comments that imply negative sentiment. Similar inclination was observed for these steps in all three runs of the MOOC. Since these two steps have shown a different trend than other steps, the reason behind this high or low scores needs to be identified. WordCloud is used to extract the most

frequent words from these comments to explore the possible explanation of these scores.

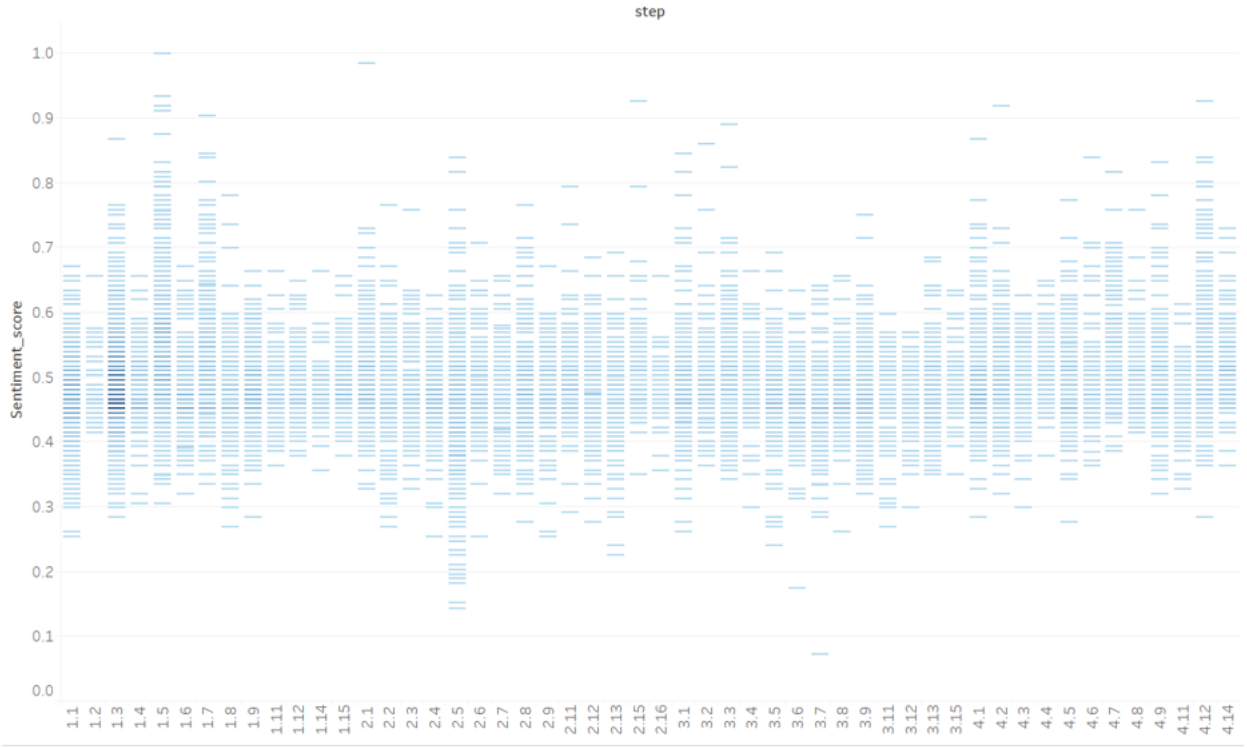

Figure 3.7: Step-wise sentiment Score for each comment

3.8 Summary

This chapter summed up all the methods that are used in the implementation of this project. It also covered the design diagrams and all the necessary formulas that are applied. An overview of all the preprocessing tasks as well as sentiment analysis method is discussed. To get an idea on the course data, step-wise and weekly analysis is performed based on visualizations. Also, to investigate the difference between the trends of the data based on their content type, the data is analyzed based on its content and the same analysis is repeated. Finally, the correlation methods are elaborated for exploratory data analysis.

Chapter 4

Results

This section aims to describe the results of the analysis and interpretation. The interpretation of the results is given based on the posed research questions. The outcomes are discussed in detail and also represented visually to gain a better understanding of them. The methodologies that were used for this analysis were Python NLTK for data preprocessing, SentiWordNet for sentiment analysis, and Pandas Profiling for exploring mutual associations and the results for each one of them is described step by step.

4.1 Data Preprocessing

The first part of an exploratory was to prepare the available data for analysis by the means of normalization techniques after tokenization and part-of-speech tagging. The original comment is read as:
[I'm from the United States. I grew up in the south and I believe our literature is a great resource.]
After tokenizing and tagging parts of speech, the text is divided into tokens as follows:
[('from', 'IN'), ('the', 'DT'), ('united', 'JJ'), ('states', 'NNS'), ('grow', 'VBP'), ('the', 'DT'), ('south', 'NN'), ('and', 'CC'), ('believe', 'VB'), ('our', 'PRP$'), ('literature', 'NN'), ('great', 'JJ'), ('resource', 'NN')]

4.2 Sentiment Analysis

As can be seen in the plot 3.7, the sentiments of learners are quite evenly distributed for all steps. Most of the comments have a neutral sentiment towards the course contents. To get an overall trend of this analysis, the average score for each comment is calculated.

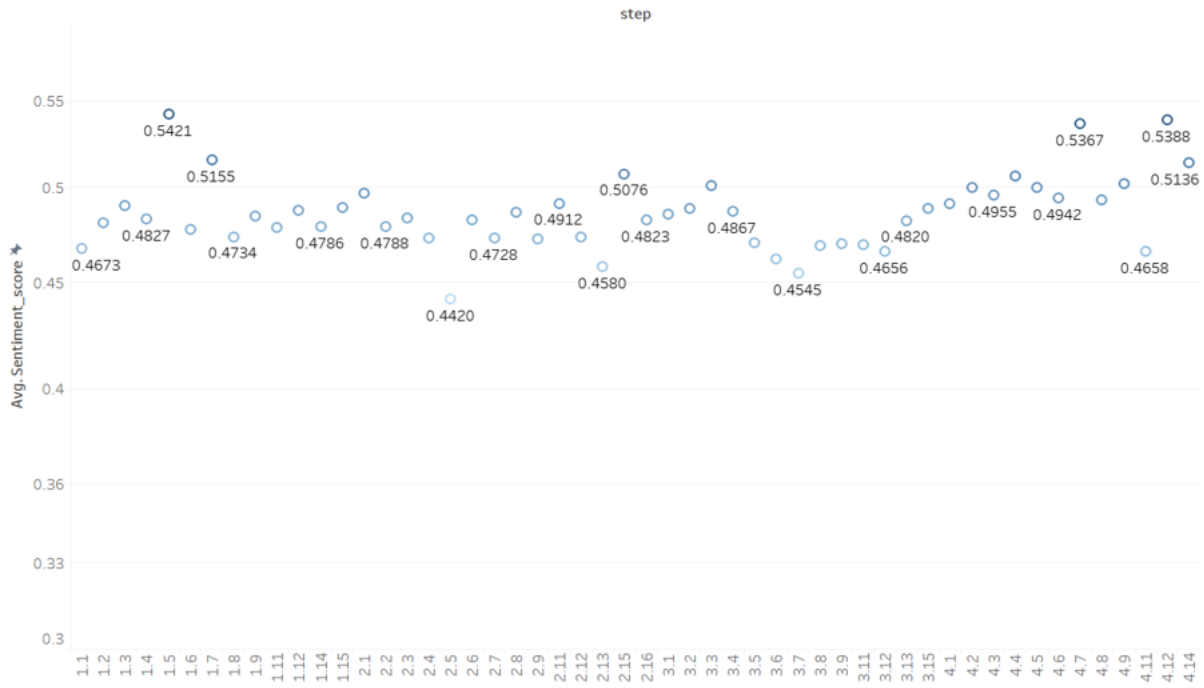

Figure 4.1: Step-wise average sentiment Score

Figure 4.1 shows the distribution of the average sentiment score of each step for MOOC run 3. The average sentiment is in the neutral range for all the steps and this indicates that the number of positive comments is almost equal to the number of negative comments. With the completion of calculation of average sentiment scores, to explore its mutual association with step completion percentage and dropouts, a deep dive into course attendance and completion rate is required.

4.3 Course Attendance and Dropouts

The number of students attending each step can be determined with the help of the activity data. The line graph below shows the total number of students that attended that particular step. As the course progresses, the gradual decrease in total attendance for the steps can be noted from the graph. The start of every week, i.e., 1.1, 2.1, 3.1 and 4.1 saw a surge in attendance from its previous step. This might be because of the learner checking the course contents before investing their time in the next steps. And therefore, these four steps were excluded from the next stage of investigating the mutual association as they might affect the correlation because of the rise in attendance. The analysis below is based on the data collected for MOOC run 3.

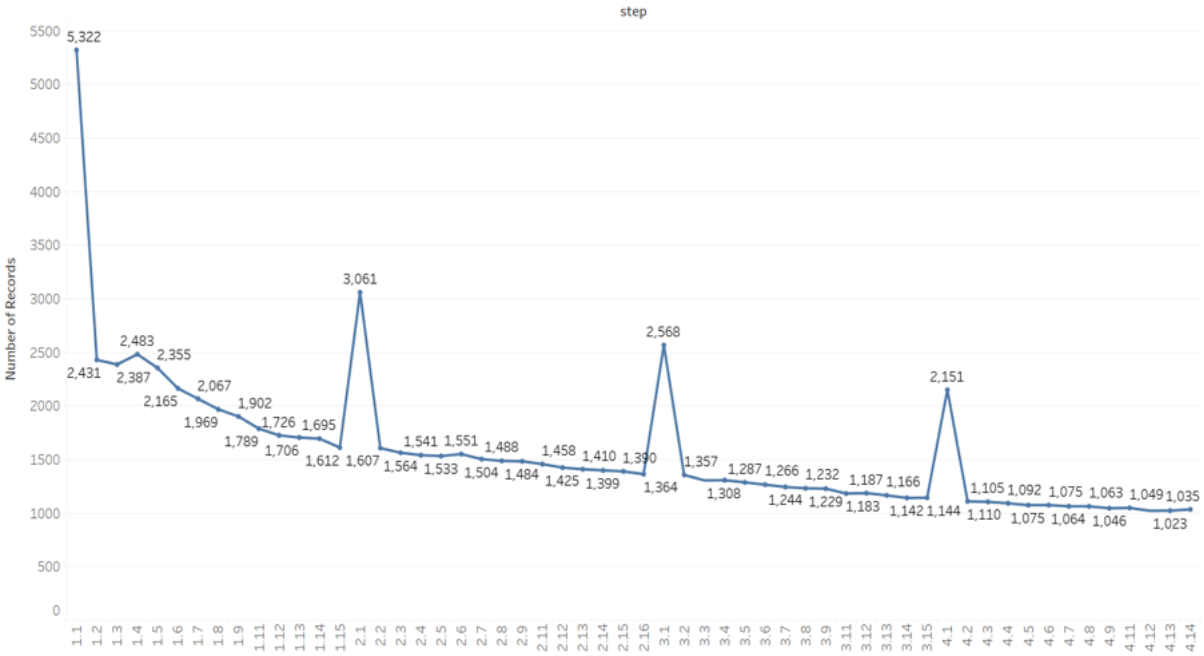

Figure 4.2: Total attendance for each step

Before proceeding to the next stage, the dropouts rate has to be calculated for all the steps. As the first steps of every week are excluded, the dropouts for the first step of week 2, 3 and 4 are considered from the last step of the previous week. For example, now the first step of the second week is 2.11 and its dropout is calculated based on the last step of the first week, i.e., step 1.9.

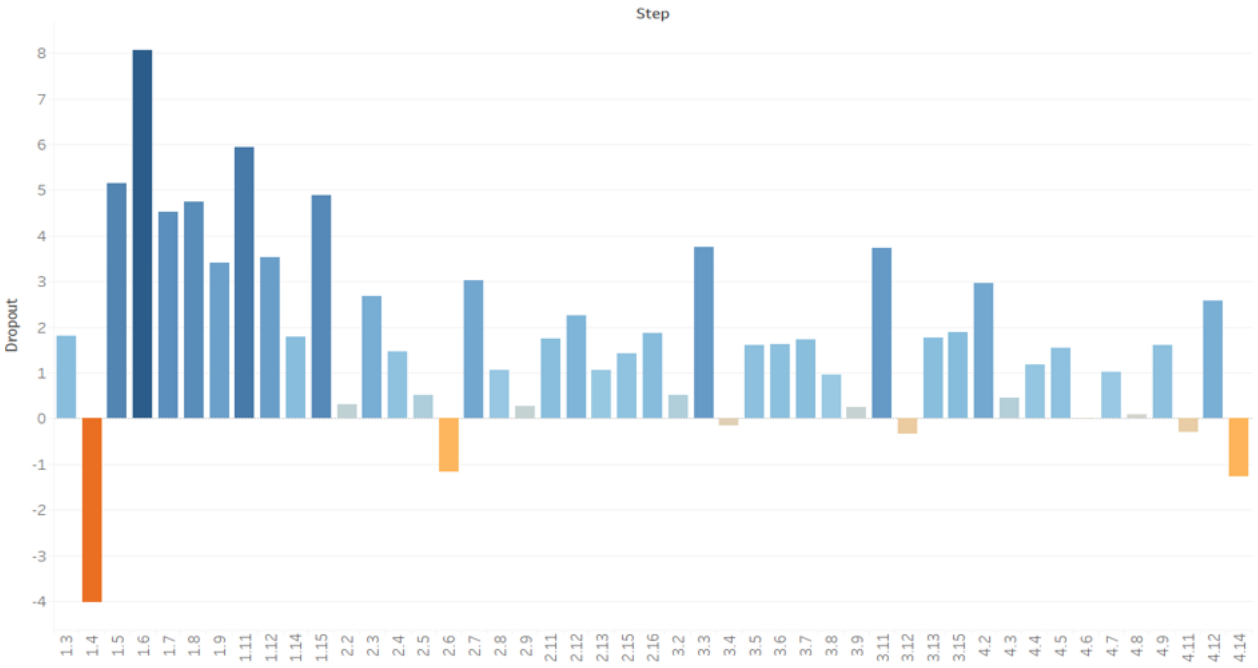

Figure 4.3: Step-wise Dropout rate

Figure 4.3 presents the details on the dropout rates for each step where positive value shows the percentage drop in attendance compared to previous steps whereas the negative value show rises in the number of audiences from the previous step. It is worthwhile noting that there is no clear trend in dropouts in terms of steps or weeks.

4.4 Correlation Outcomes

As all the calculations and analysis has been carried out, the data is ready to explore the relation of sentiments of the authors of comments with step completion percentage and the dropout rate. As the sentiment scores are between 0 to 1, the other values, i.e., percentage and dropouts are also converted to a range of 0 to 1. The pandas profiling module gives an HTML report after passing the variables that include the correlation matrix and other important statistics.

Firstly, when the step completion percentage and average sentiment scores for each step are passed as an input, the pandas profiling did not find any mutual association between the two for any of the runs of MOOC. All four correlation coefficients, Pear-

son's r, Spearman's ρ, Kendall's τ, and Phik ϕk, gave results close to 0 which signifies no correlation between step completion percentage and average sentiment score. The correlation matrix for these two variables for MOOC run 3 data is given below. The matrices above clearly suggest that there is no correlation between the step completion and sentiments.

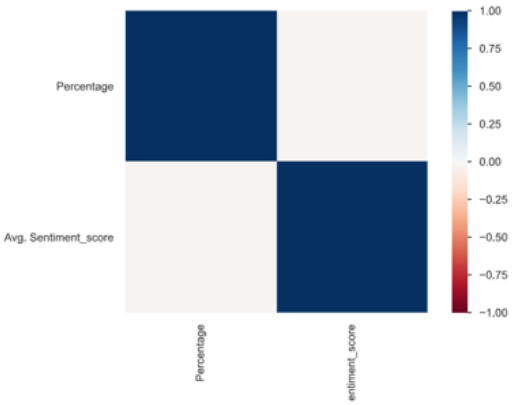

Figure 4.4a Person's r

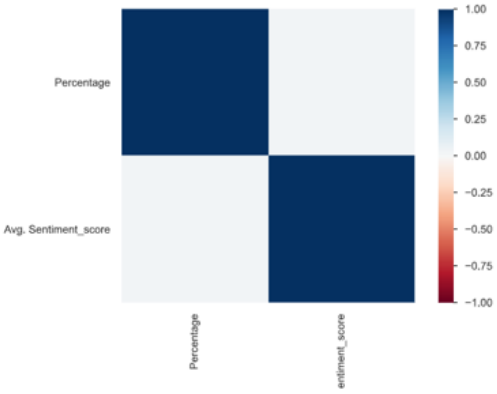

Figure 4.4b Spearman's ρ

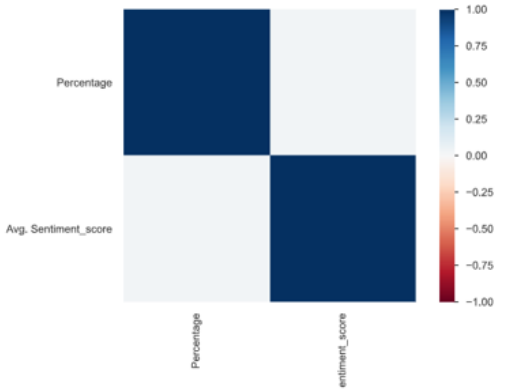

Figure 4.4c Kendall's τ

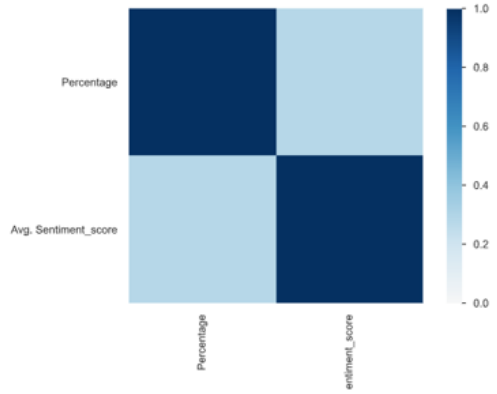

Figure 4.4d Phik φk

Figure 4.4: Coefficient matrices for step completion rate and sentiment scores

The exact values for these coefficients for each run are given below. These correlation values from table 1 prove that there is no linear, ordinal, or non-linear relationship between step completion percentage and the average sentiment value of comments on that step.

Correlation Coefficient	Run 1	Run 2	Run 3
Pearson's correlation	-0.035	-0.126	-0.024
Spearman's correlation	-0.114	-0.007	0.028
Kendall's correlation	0.010	0.008	0.029
Phik correlation	0.006	0.003	0.211

Table 4.1: Correlation Coefficients for step completion

Coming to the next stage, the dropout values at the particular step and average sentiment score from the previous step comments are given as input to the pandas profiling module. Another report is generated in HTML format that gives the values of correlation coefficients and other statistics. The correlation matrices were as follows:

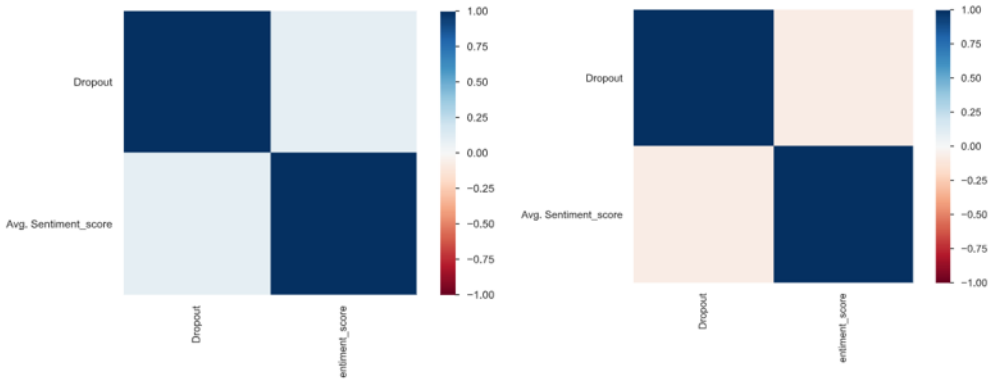

Figure 4.5a Person's r

Figure 4.5b Spearman's ρ

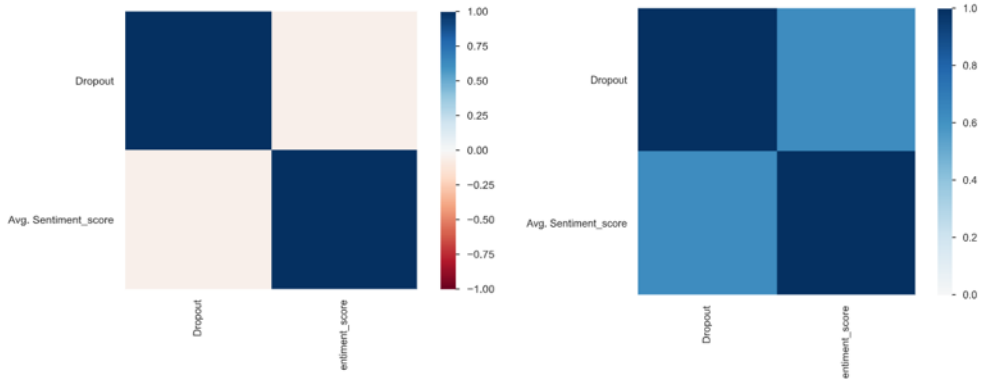

Figure 4.5c Kendall's τ

Figure 4.5d Phik φk

Figure 4.5: Coefficient matrices for dropouts and sentiment scores

The above correlation matrices give the strength of association between the attrition rate and the average sentiment value of the previous step. Although the values for Pearson's r, Spearman's ρ, and Kendall's τ do not show any sign of dependency between the two variables, Phik ϕk has a value more than 0.5 that shows some association between the two variables. The values of the coefficients are –

Correlation Coefficient	Run 1	Run 2	Run 3
Pearson's correlation	-0.108	-0.137	-0.087
Spearman's correlation	-0.172	-0.186	-0.075
Kendall's correlation	-0.004	-0.119	-0.049
Phik correlation	0.311	0.330	0.620

Table 4.2: Correlation Coefficients for dropouts

From the first three coefficients, it can be understood that there is no strong correlation between the dropouts and sentiment scores. The values for Pearson coefficient were too low to justify any linear relationship and therefore, the Phik coefficient which gives the value of 0.62 for run 3, indicates towards non-linearity between the two variables but it is not significant enough to prove the variables have a complete non-linear dependency. The two variables are plotted against each other to check if they have any dependency between them. If the two variables do have some consistency with each other even with different ratios, it will be visible through the plot.

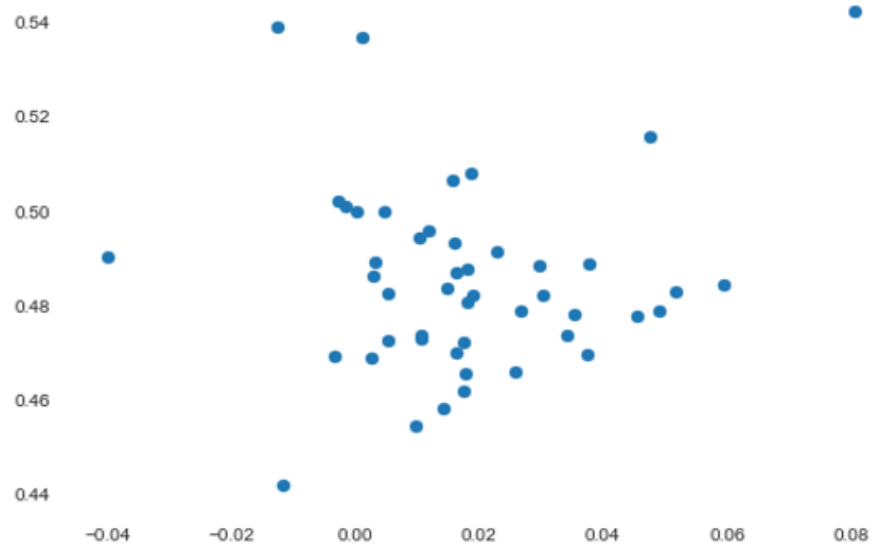

Figure 4.6: Dropouts vs Average Sentiment Score

The above diagram is non-consistent and does not show any sign of non-linear relation. An additional test of non-linear dependency was done by using 'nlcor' package in R that gave the result as follows:

cor.estimate	0.2773882

Table 4.3: Non-linear Correlation Coefficients for Run 3

This correlation value is less than 0.3 which confirms that the non-linear relation between the dropouts and average sentiments of students is not substantial.

To summarise, after preparing the data for analysis, the strength of association of dropouts and completion percentage with average sentiment scores is calculated. With the use of four correlation methods, it was proved that there is no dependency between any of the variables. Even though Phik coefficient showed some non-linear dependency between dropout and sentiment score for one of the runs, it was minimal and hence additional tests were performed. The additional tests also were not significant in proving any dependency between the variables.

4.5 Content-wise analysis

The next part of the analysis was to calculate the association of the sentiments with dropouts and step completion rate for each of the content types. The analysis shows that the average sentiment scores of articles were greater than that of the videos. In addition to this, it was clearly noticeable from figure 3.6 the student attendance for video steps also fell at a higher rate than article steps. In the next stage, the sentiment scores along with step completion rate were given as an input to the pandas profiling module in python, just similar to the analysis of whole data.

The HTML report gives the correlation coefficient values for both articles and videos one by one separately. The coefficient matrices generated for videos and articles for MOOC run 3 are given as:

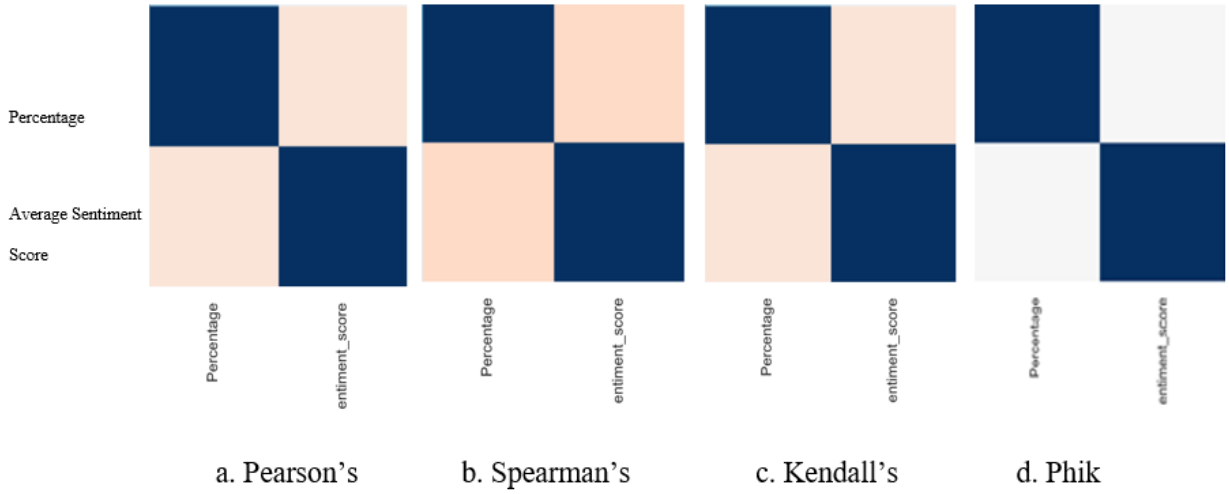

Figure 4.7: Correlation matrices for article step completion

Correlation Coefficient	Run 1	Run 2	Run 3
Pearson's correlation	0.416	0.288	0.39
Spearman's correlation	0.437	0.361	0.55
Kendall's correlation	0.319	0.200	0.38
Phik correlation	0.465	0.680	0.54

Table 4.4: Correlation Coefficients for videos - Step completion

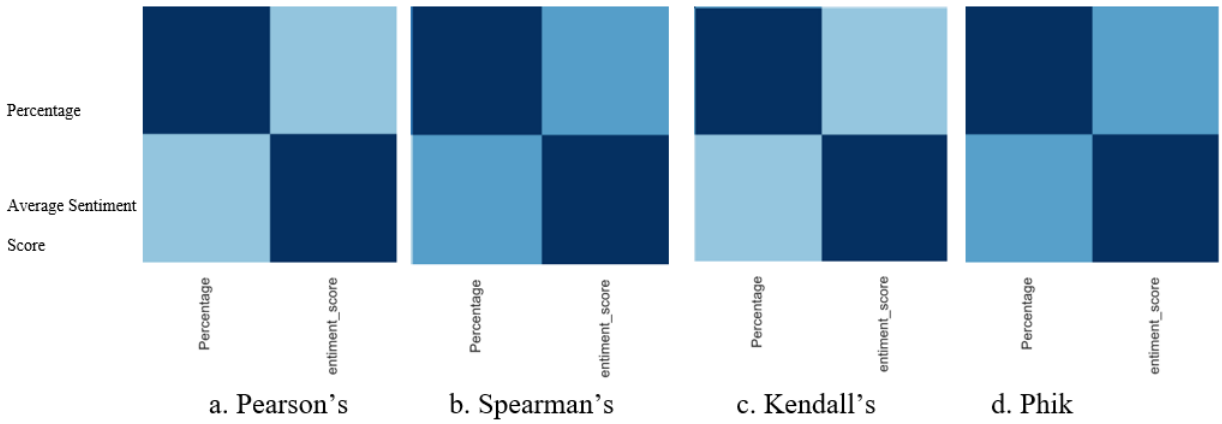

Figure 4.8: Correlation matrices for video step completion

Correlation Coefficient	Run 1	Run 2	Run 3
Pearson's correlation	-0.317	-0.280	-0.235
Spearman's correlation	-0.312	-0.065	-0.197
Kendall's correlation	-0.214	-0.025	-0.131
Phik correlation	0.25	0	0

Table 4.5: Correlation Coefficients for Articles - Step completion

As can be seen from the tables 4.4 and 4.5, the coefficients for videos are around the 0.5 marks for all three runs which depict that there is some correlation between the step completion ratio and comments on that particular step. All the coefficients give a significant score based on which we can conclude that the step completion ratio falls in many cases where the sentiments are negative and vice versa. Hence, this result proves that if there are many students who leave the step incomplete, the comments have a negative average score. To confirm this result, the video step completion rate was plotted against the average sentiment scores for that step. The below plot shows that there is some linearity between the two variables and they do have some linear association with each other.

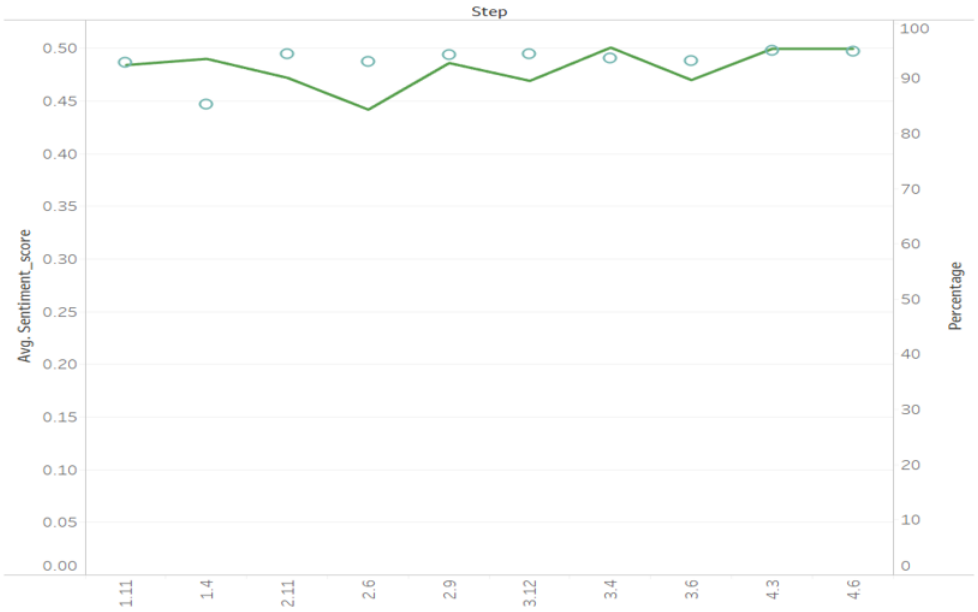

Figure 4.9: Video step completion rate vs average sentiment score

On the other hand, for the articles, the coefficient values do not show any correlation at all as the values are about zero. It is enough to say that there is no association at all between step completion rate and sentiments for article steps.

Following step completion rates, the dropout rate of the step with sentiment scores of the previous step is provided as an input to the pandas profiling. The output in the form of a report gives the following matrices for MOOC run 3:

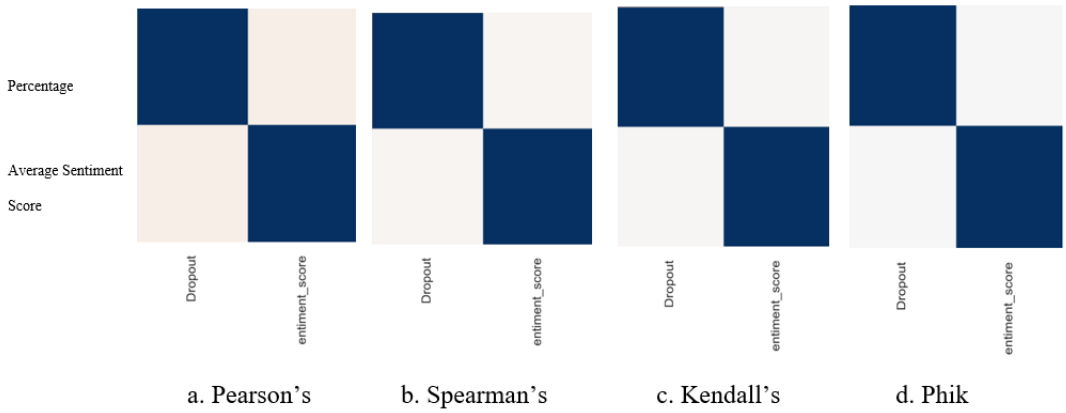

Figure 4.10: Correlation matrices for article step dropouts

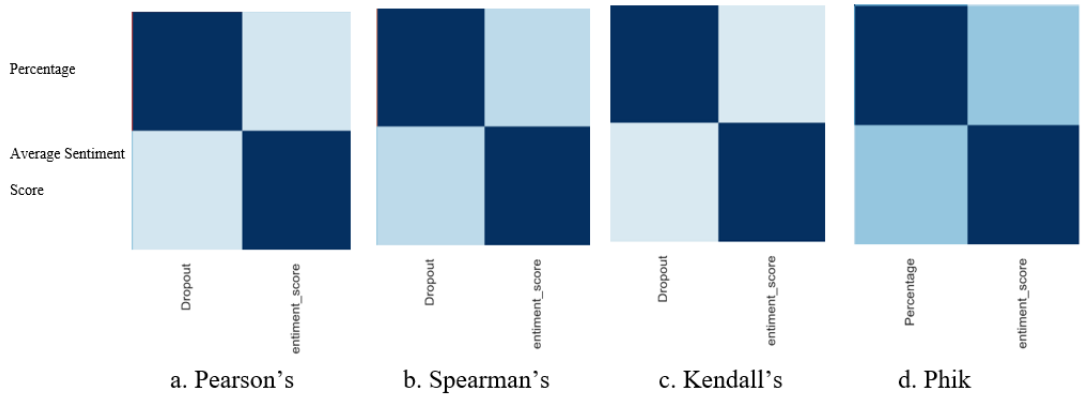

Figure 4.11: Correlation matrices for video step dropouts

Below tables elaborates the exact values for the coefficients:

Correlation Coefficient	Run 1	Run 2	Run 3
Pearson's correlation	-0.316	-0.018	0.181
Spearman's correlation	-0.139	-0.117	0.162
Kendall's correlation	-0.178	-0.085	0.156
Phik correlation	0.260	0.314	0.434

Table 4.6: Correlation Coefficients for videos - Dropouts

Correlation Coefficient	Run 1	Run 2	Run 3
Pearson's correlation	-0.132	-0.149	-0.055
Spearman's correlation	-0.164	-0.248	-0.020
Kendall's correlation	-0.099	0.159	-0.011
Phik correlation	0.215	0.250	0.207

Table 4.7: Correlation Coefficients for Articles - Dropouts

The values for both videos and articles for all three runs do not show any association between the dropouts and the sentiment scores of comments. The values are not consistent and they do not have similar trends in any of the runs to establish any kind of dependency. This proves that the sentiments of the students do not play any role in dropouts for any of the specific content type or the overall course.

4.6 Step analysis (1.5 & 2.5)

As it was seen that steps 1.5 and 2.5 have a dissimilar trend to other comments from the dataset, it is important to analyze the reason behind this difference. To examine this, the comments of these two steps are analyzed with the help of WordCloud and Feature Extraction libraries of Python. These libraries will help in identifying the particular reasons behind the inclination towards positive or negative scores for these steps.

After completing all the preprocessing tasks and selecting the comments on the steps 1.5 and 2.5, the most used words, i.e, frequency of these words in the comments is calculated. The analysis for two steps is done separately in order to obtain two different word clouds. Using these words, the word cloud is formed and based on the figure, it is possible to get an idea of the highly frequent words that affect the sentiment score in either direction.

Figure 4.12: Word Cloud for Step 1.5

World cloud in figure 4.12 shows the words with highest frequency in the forum posts of step 1.5. The large font denotes that the word is used more than others and as the font goes smaller, the word is less frequent. It is visible from this word cloud that there are many positive words such as god, church, beautiful, monk, gospel, bible, devotion, like and many more. These words do contribute to the overall sentiments of the sentence which makes the overall score is higher. This was found to be the reason behind the

step 1.5 having more positive comments than other steps.

Coming to the word cloud of step 2.5, it can be seen from figure 4.13 that the words mentioned in the comments are oriented towards the neutral or negative side. Words such as difficult, problem, material, lack, poor bad, physical which are used in these comments give a score less than zero which leads to an overall negative score of the comment.

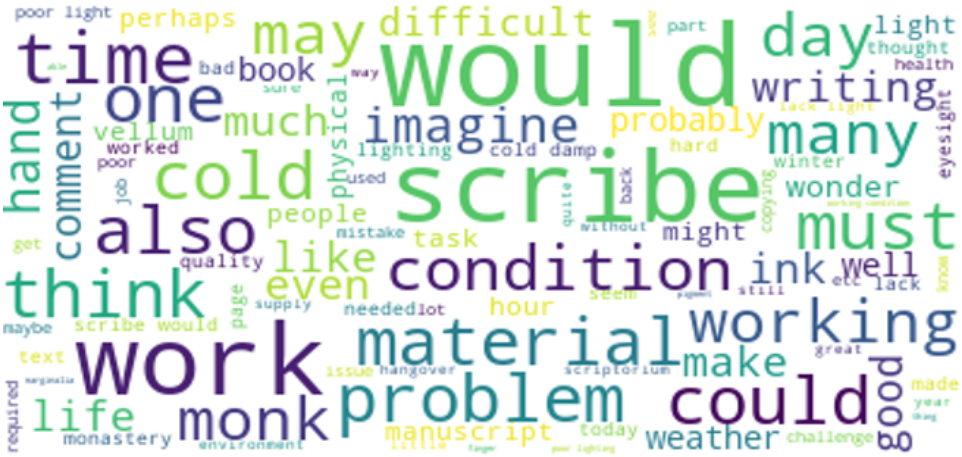

Figure 4.13: Word Cloud for Step 2.5

To explore the reasons behind these negative or positive words and their presence in all three runs of the MOOC, the topic of the course for these steps was examined. The type of content for step 1.5 is an article based on the creators of the Book of Kells, i.e, Monks and step 2.5 is a discussion step asking a question about the problems that occurred in their work. As the article in 1.5 is based on the creators, it includes information about Cristian religion and their God. Based on the content, i.e, the topic of this step, the comments are also revolved around the same category and include a lot of positive words. As a result, this topic leads to high sentiment score for the comments.

On the contrary, step 2.5 has comments discussing the problems encountered while working or the conditions in which the scribe worked. Therefore the comments have a lot of words related to the difficulties and as a result, the sentiment score for these comments is low. Similar to the above case, the topic of this step also influences the overall sentiment score.

4.7 Observations

When the analysis was done on the whole data, the results suggest that there is no mutual association between the dropouts or step completion rate with the average sentiment score for the Book of Kells MOOC data. All three runs show similar trends in all the coefficient values and all these values are close to zero which fails to prove that there is any correlation at all. Through the content-wise analysis, it can be depicted that the video step completion rate has a moderate correlation with the average sentiment scores of the step. Conversely, for article steps, no correlation was found in step completion rates and the sentiments. Coming to the analysis of dropouts with respect to the content types, no correlation was found with the average sentiment scores.
Based on these observations, it is possible to say that only one of the four hypotheses holds.

• Hypothesis 1: Even though there was a little significance in the non-linear relationship between dropouts rate and sentiment scores, based on other results it can be deduced that there is no relationship between the two and hypothesis fails to hold.

• Hypothesis 2: The coefficients are too low for sentiment scores and step completion rate for the whole data and subsequently, the second hypothesis does not hold.

• Hypothesis 3: Analyzing the data based on its content types showed that there is no difference in correlation values of attrition rate and sentiment score in videos or articles. Both content types had association values close to zero similar to the full data analysis. Therefore, the hypothesis does not hold.

• Hypothesis 4: The correlation coefficient for step completion rate and sentiment scores for videos and articles did vary by a good margin. Therefore, the last hypothesis holds as step completion rate and sentiment score for videos do have an adequate correlation whereas there is no association between the two for articles. Furthermore, steps 1.5 and 2.5 observed very positive and very negative scores respectively, the main reason was identified as the topic on which the MOOC step was based.

Chapter 5

Discussion

This section explains the possible reasons behind the results of this analysis and also draws a comparison between the results of this thesis with the previous findings. This analysis was based on the data of a MOOC, available on the FutureLearn website. After calculating the sentiment scores from the comments on the forum and dropouts rate, their correlation was identified and the outcome of this analysis suggest that there is no strength of association between the students' sentiments and the attrition rate or step completion rate. Furthermore, when the steps are divided according to their content type, only video steps have some amount of dependency on their completion rates with the sentiment scores.

Several aspects of the MOOC are considered while exploring the possible explanations of these results. The first and most important attribute assessed was the ratio of the number of students who attend the course to the number of students who comment on the discussion forums. For example, if we consider MOOC run 1, there are 8701 distinct users who attend the course whereas only 3083 of them comment on the forum during the course time. As the sentiment analysis was performed on the comments, it is possible to extract the opinions of these 3083 students. There is a huge difference between these numbers and therefore, there is a possibility that this might be one of the important reasons behind the lack of correlation between student sentiments and dropouts.

While most of the previous research of predicting the dropouts rate does not associate the attrition rate directly with sentiment analysis, Chaplot et al. [2015] in their

research consider student sentiments as an important attribute. One of the main limitations of the research done by Chaplot et al. [2015] can be considered as they do not explore the relationship between sentiments with dropouts before applying it to the prediction model along with other attributes. In order to get more active participation in MOOCs and increase the number of comments, Adamopoulos [2013] along with his research team gets enrolled in MOOCs. This results in richer texts that give better results for their predictive model.

Another possible reason behind the outcome of this analysis can be the topic on which the comments are based. As the words used in these comments can be related to a particular topic which may have an impact on the overall sentiment score of that comment. A perfect example of this was seen in the stepwise analysis to check if any of the steps do have extremely positive or extremely negative scores. While measuring student performance, Alblawi & Alhamed [2017] adds sentiment scores for student's feedback rather than using student's comments. This ensures that the calculated sentiment score is about the course and does not have any bias.

Chapter 6

Conclusion and Future Work

The goal of this chapter is to conclude the thesis by discussing all the processes that were undertaken during the analysis and the implications of their results. In addition to this, limitations of the thesis and future works are described in this chapter.

6.1 Conclusion

The foremost goal of this dissertation was to explore the relation between the attrition rate in MOOCs and the student sentiments obtained through posts on the forum. Furthermore, the step completion ratio was also compared with the student sentiments to study their relation. This analysis was performed on the MOOC based on Book of Kells and same analysis was repeated after dividing the data according to the step type such as Videos or Articles.

After calculating the attrition and step completion rates and performing data preprocessing tasks such as tokenization, POS tagging, and Lemmatization, the sentiments of the learners were extracted from the comments.Python NLTK and SentiWordNet were used for these preprocessing and sentiment analysis tasks.Exploring correlation between the two was the next task and that was accomplished with the help of pandas profiling module in python. It measures the four coefficients for the correlations namely, Pearson's r, Spearman's ρ, Kendall's τ, and Phik ϕk.

On the basis of data considered, the conclusion of the analysis can be drawn as the sentiments of the students do not affect the dropouts rate in the MOOC and it is

consistent across all content types of it. The ratio of number of students who attend the course to the number of students who interact on forum is high and that might have led the analysis to this outcome. When the data was divided according to their step types, it was observed that the step completion ratio for all other content types other than videos has no correlation with student sentiments.Digging into the comments and their sentiment scores, it was perceived that most of the times, the sentiment score depends upon the type of topic on which the MOOC step is based as the words used in the comments are related to the topic.

6.1.1 Key Findings

- Question 1
  - Considering the full data, sentiments of student comments do not have any correlation with the attrition rate or the step completion rate.

    Data was prepared for the analysis with the help of Natural Language Processing in python. Selection of method for sentiment analysis was based on the output format of the sentiment scores from three well-known methods.

- Question 2
  - Considering the content-wise data, the correlation values are consistent with the attrition rate but do vary for step completion rate.
    1. Sentiments of the student comments did not differ extremely across all content types.
    2. The sentiment scores for comments on two steps do have extreme values that are dissimilar to other steps. The topic content of these steps is proved to be the reason behind this difference.

6.1.2 Limitations

As it can be seen from the outcome of the analysis of sentiment scores that the words which are related to the topic of the MOOC such as "Problem" or "Church" are considered to be very positive or very negative, and this misleads the overall score of the sentence. This misleading score can sometimes fail to show the truthfulness of the student's sentiment.

The data considered for this analysis was had three different runs from the same MOOC but the analysis may be give different results if data for other MOOCs is considered. The topic of the MOOC can influence the comments and subsequently, it does impact the sentiment score with which the attrition rate is compared.

6.2 Future Work

The student sentiments can be influenced by the words used related to the topic of the MOOC rather than actual content.For example, as this analysis was based on a MOOC that explains history, it had most of the comments related to the problems they faced or related to religion in that era. To avoid this, sentiment scores of the comments can be combined with a survey filled by all learners that describe their attitude towards the course. This survey should contain all the questions related to the feedback for the course undertaken, so that student sentiments are accurate.

In order to avoid bias based on the topic of a step, the words related to the topic can be excluded while calculating the sentiment scores if it is not possible to add a mandatory survey. Also, according to Balch [2013], commitment level of the students for MOOCs may differ from the regular courses which may result into high attrition rate and therefore, the reasons behind the dropouts can be considered.

Bibliography

[Adamopoulos, 2013] Adamopoulos, P. (2013). What makes a great mooc? an interdisciplinary analysis of student retention in online courses. *Thirty Fourth International Conference on Information Systems*.

[Alblawi and Alhamed, 2017] Alblawi, A. S. and Alhamed, A. A. (2017). Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, nlp and analytics. *IEEE Conference on Big Data and Analytics (ICBDA)*.

[Amrrs, 2018] Amrrs (2018). https://www.kaggle.com/nulldata/intro-to-pandas-profiling-simple-fast-eda.

[Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining‘. *Lrec*.

[Balch, 2013] Balch (2013). https://augmentedtrader.com/2013/07/24/why-the-low-mooc-completion-rate-statistic-is-a-bogus-argument/.

[Baturay, 2015] Baturay, M. H. (2015). An overview of the world of moocs. *Procedia - Social and Behavioral Sciences*, pages 427–433.

[Brugman, 2020] Brugman (2020). https://github.com/pandas-profiling/pandas-profiling.

[Chaplot et al., 2015] Chaplot, Rhim, E., and Kim, J. (2015). Predicting student attrition in moocs using sentiment analysis and neural networks. *AIED Workshops*, pages 54–57.

[Chen and Sokolova, 2018] Chen, Q. and Sokolova, M. (2018). Word2vec and doc2vec in unsupervised sentiment analysis of clinical discharge summaries.

[Chen and Zhang, 2017] Chen, Y. and Zhang, M. (2017). Mooc student dropout: pattern and prevention. *Proceedings of the ACM Turing 50th Celebration Conference*.

[Cobos et al., 2019] Cobos, R., Jurado, F., and Blázquez-Herranz, A. (2019). A content analysis system that supports sentiment analysis for subjectivity and polarity detection in online courses. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*.

[Coffrin et al., 2014] Coffrin, C., Corrin, L., de Barba, P., and Kennedy, G. (2014). Visualizing patterns of student engagement and performance in moocs. *Proceedings of the fourth international conference on learning analytics and knowledge*, pages 83–92.

[Deshpande and Chukhlomin, 2017] Deshpande, A. and Chukhlomin, V. (2017). What makes a good mooc: A field study of factors impacting student motivation to learn. *American Journal of Distance Education*, pages 275–293.

[Estrada et al., 2020] Estrada, M. L. B., Cabada, R. Z., Bustillos, R. O., and Graff, M. (2020). Opinion mining and emotion recognition applied to learning environments. *Expert Systems with Applications*.

[Esuli and Sebastiani, 2006] Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

[Ferschke et al., 2015] Ferschke, O., Yang, D., Tomar, G., and Rosé, C. P. (2015). Positive impact of collaborative chat participation in an edx mooc. *International Conference on Artificial Intelligence in Education*, pages 115–124.

[Fournier and Kop, 2015] Fournier, H. and Kop, R. (2015). Mooc learning experience design: issues and challenges. *International Journal on E-Learning*, pages 289–304.

[FutureLearn, 2012] FutureLearn (2012). https://www.futurelearn.com/.

[Harris et al., 2014] Harris, S. C., Zheng, L., and Kumar, V. (2014). Multi-dimensional sentiment classification in online learning environment. *IEEE Sixth International Conference on Technology for Education*, pages 172–175.

[Hew et al., 2020] Hew, Foon, K., Hu, X., Qiao, C., and Tang, Y. (2020). What predicts student satisfaction with moocs: a gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers Education*.

[Hew and Cheung, 2014] Hew, K. F. and Cheung, W. S. (2014). Students' and instructors' use of massive open online courses(moocs): Motivation and challenges. *Educational research review*, pages 45–58.

[Iraj, 2014] Iraj (2014). https://www.linkedin.com/pulse/mooc-completion-rates-what-do-you-think-hamideh-iraj/.

[Jordan, 2015] Jordan, K. (2015). Massive open online course completion rates revisited: Assessment, length and attrition. *International Review of Research in Open and Distributed Learning*, pages 341–368.

[Kagklis et al., 2015] Kagklis, V., Karatrantou, A., Tantoula, M., Panagiotakopoulos, C. T., and Verykios, V. S. (2015). A learning analytics methodology for detecting sentiment in student fora: A case study in distance education. *European Journal of Open, Distance and E-learning*.

[Keegan, 2020] Keegan (2020). https://skillscouter.com/online-learning-statistics/.

[Kumar and Jain, 2015] Kumar, A. and Jain, R. (2015). Sentiment analysis and feedback evaluation. *IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)*, pages 433–436.

[Liberman, 2003] Liberman, M. (2003). Penn treebank p.o.s. tags.

[Liu et al., 2016] Liu, Z., Liu, S., Liu, L., Sun, J., Peng, X., and Wang, T. (2016). Sentiment recognition of online course reviews using multi-swarm optimization-based selected features. *Neurocomputing*, pages 11–20.

[Marta, 2019] Marta (2019). https://brand24.com/blog/the-benefits-of-sentiment-analysis/.

[Maynard and Funk, 2011] Maynard, D. and Funk, A. (2011). Automatic detection of political opinions in tweets. In *Extended Semantic Web Conference*, pages 88–99. Springer.

[Min et al., 2019] Min, L., Wenting, Z., Shi, Y., Pan, Z., and Li, C. (2019). What do participants think of today's moocs: an updated look at the benefits and challenges of moocs designed for working professionals. *Journal of Computing in Higher Education*.

[Mora, 2012] Mora, S. L. (2012). Brief (very brief) history of moocs.

[Moreno-Marcos et al., 2018a] Moreno-Marcos, P., Alario-Hoyos, C., PJ, M.-M., Estévez-Ayres, I., and CD, K. (2018a). A learning analytics methodology for understanding social interactions in moocs. *IEEE Transactions on Learning Technologies.*, pages 442–455.

[Moreno-Marcos et al., 2018b] Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., Estévez-Ayres, I., and Kloos, C. D. (2018b). Sentiment analysis in moocs: A case study. *IEEE Global Engineering Education Conference (EDUCON)*.

[Peng and Xu, 2020] Peng, X. and Xu, Q. (2020). Investigating learners' behaviors and discourse content in mooc course reviews. *Computers Education*.

[Raschka, 2014] Raschka, S. (2014). https://sebastianraschka.com/articles/2014_about-_feature_scaling.html.

[Reich, 2014] Reich, J. (2014). https://er.educause.edu/articles/2014/12/mooc-completion-and-retention-in-the-context-of-student-intent.

[Shah, 2019] Shah, D. (2019). https://www.classcentral.com/report/mooc-stats-2019/.

[Shapiro et al., 2017] Shapiro, H. B., Lee, C. H., Roth, N. E. W., Li, K., Çetinkaya Rundel, M., and Canelas, D. A. (2017). Understanding the massive open online course (mooc) student experience: An examination of attitudes, motivations, and barriers. *Computers Education*.

[Sharkey and Sanders, 2014] Sharkey, M. and Sanders, R. (2014). A process for predicting mooc attrition. *EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 50–54.

[Singh, 2019] Singh, V. (2019). https://medium.com/@silentflame/pearson-correlation-a-mathematical-understanding-c9aa686113cb.

[SocialLearning, 2015] SocialLearning (2015). http://sociallearningcommunity.com/10-of-the-best-mooc-providers/.

[Sunar et al., 2017] Sunar, A. S., White, S., Abdullah, N. A., and Davis, H. C. (2017). How learners' interactions sustain engagement: A mooc case study. *IEEE Transactions on Learning Technologies*, pages 475–487.

[Suzuki et al., 2006] Suzuki, Y., Takamura, H., and Okumura, M. (2006). Application of semi-supervised learning to evaluative expression classification. *International Conference on Intelligent Text Processing and Computational Linguistics*.

[Wang et al., 2018] Wang, L., Hu, G., and Zhou, T. (2018). Semantic analysis of learners' emotional tendencies on online mooc education. *Sustainability*.

[Wen et al., 2014] Wen, M., Yang, D., and Rose, C. (2014). Sentiment analysis in mooc discussion forums: What does it tell us? *Educational data mining*.

[Widyahastuti and Tjhin, 2018] Widyahastuti, F. and Tjhin, V. U. (2018). Performance prediction in online discussion forum: state-of-the-art and comparative analysis. *Procedia Computer Science*, pages 302–314.

[Xing, 2019] Xing, W., . D. D. (2019). Dropout prediction in moocs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, pages 547–570.

[Yin, 2019] Yin, S. (2019). The analysis and early warning of student loss in mooc course. *ACM Turing Celebration Conference - China (ACM TURC 2019)*.

[Yu, 2010] Yu, C. H. (2010). Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research*.

[Zhang et al., 2018] Zhang, C., Chen, H., and Phang, C. W. (2018). Role of instructors' forum interactions with students in promoting mooc continuance. *Journal of Global Information Management*, pages 105–120.

Appendix

Appendix I - POS Tags

Please see Table 1 on page 56 for description of the POS tags used in this dissertation.

Sr No	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Table 1: POS tags used [Liberman, 2003]