THE UNIVERSITY
of ADELAIDE

Investigating Students' Learning through Learner roles and Linguistic Expressions in Massive Open Online Courses

by
Lavendini Sivaneasharajah

Supervised by:
Professor Katrina Falkner
Dr Thushari Atapattu
Dr Rebecca Vivian

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the
Faculty of Engineering, Computer and Mathematical Sciences
School of Computer Science
THE UNIVERSITY OF ADELAIDE

May 2022

Abstract

In recent years, there has been increasing interest in online learning environments, notably in Massive Open Online Courses (MOOCs). Due to such interest, predictions and education data mining have rapidly gained prominence in education studies over the past decade. As many of the MOOCs are freely available for students, they draw the interest of thousands of learners. However, assessing the *success rate* of student learning through online platforms has become difficult to quantify as students enroll for varying purposes, such as browsing the course content or enrolling into similar courses to find the best fit for their needs. Knowing that students may enroll in courses for other purposes, research studies need to explore diverse perspectives of learning success beyond *completion*. The massive amount of student data available in MOOC platforms enables researchers to gain valuable insights into students' learning behaviours, enabling analysis of aspects such as performance predictions and cognitive engagement. Using discourse analysis, it is possible to investigate the learner-generated discourse from discussion forums to understand students' learning in many ways so as to identify information-seeking learners, to identify linguistic behaviours of students working at different grades and to understand how well a student has understood course content; in particular, their issues and knowledge on a specific content across the course.

This thesis explores the concepts of 'student roles' and 'linguistic expressions', which can be extracted from discussion forums to investigate students' learning across AdelaideX MOOCs. Using the grounded theory approach, the thesis identifies student roles in the study data as 'information-seeker', 'information-giver' and 'other'. By identifying these roles, it became possible to determine the student roles in discussion forums where a lack of peer interactions is observed. This thesis aims to categorise these roles solely based on discourse analysis by leaving the contextual (e.g., previous post) and community-related features (e.g., views) behind. The existing literature has also identified a number of important community-related structural features, such as structural position in thread and number of votes; however, a challenge is that they are not feasible to incorporate in real-time predictions as they are dynamic and change throughout the course. Furthermore, waiting for such community-related features requires more time and effort to predict these student roles in discussion forum posts. This thesis bridges this gap by predicting the student roles based solely on analysing the linguistic expressions (e.g., word count, cognitive level and analytical thinking) extracted from the content of posts.

Moreover, the thesis also identifies the learner topics that have been discussed in the forums and measures the correlation between learner clusters and topics. Exploring the

correlations, such as topic contributions with course grades, topic contributions with student roles and so forth, helps to identify how different groups of learners are contributing in discussion forums.

Going beyond student roles, this thesis presents a linguistic-based rule set to identify at-risk learners based on the linguistic contribution they have made in discussion forums which may support educators and researchers to find the associations between user-generated content and final course grades. These rule sets are generated by considering the learners' optional participation, which can be seen in many MOOCs.

Lastly, this thesis investigates the linguistic expressions of pass and fail grade learners with time for two different discussion forum components, namely comment threads and comments. Furthermore, 'Linguistic Profiles' for two different learner grades were proposed that can be used as a template to distinguish their linguistic behaviours.

Based on these investigations, the thesis provides empirical evidence of where roles exhibited by a student and their language use in discussion forums can help researchers and educators to understand the students' learning processes in an online learning environment. For example, contributions to a discussion forum by an information-giver on a discussion topic can be drastically different from students who seek information. Similarly, significant differences can be observed in the linguistic expressions exhibited by two different learner groups (pass-grade learners and fail-grade learners). The thesis also advances understanding of student learning to an extent by presenting machine learning models, topic models and decision-making rule sets that provide meaningful insights to both students and education providers in an online learning medium, especially MOOCs.

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Lavendini Sivaneasharajah

May 2022

Acknowledgements

There are numerous people who have immensely supported me throughout my PhD journey without whom this thesis would not have been possible. First and foremost, I would like to express my sincere gratitude to my supervisor Professor Katrina Falkner for her enthusiastic support, patient guidance, and the provision of outstanding facilities to work on this research project. She has also been unfailing in her encouragement and belief in my ultimate success, and extremely helpful in shaping my professional career. My deep appreciation goes to my two co-supervisors Dr Thushari Atapattu and Dr Rebecca Vivian for their immense guidance and constructive suggestions throughout this research. I also greatly admire the support provided when writing the thesis to correct them and to provide the constructive feedback.

I am grateful for the scholarship that was provided by The University of Adelaide to continue my PhD work without financial hardships and also for the Adelaide graduate research school to make the paper works easy during the submission of my thesis. I wish to thank to my research group, CSER (Computer Science Education Research) for their immense contribution and encouragement. I extend my sincere thanks to AdelaideX team for providing the research data and those who have participated for data annotation at the early stage of research study. I would like to thank Dr Alison-Jane Hunter for her professional service in proofreading my thesis. I am thankful to my fellow PhD students in our lab, for making my stay more enjoyable with lots of fun.

Last but not least, heartfelt acknowledgements are expressed to my father Sivaneasharajah and my mother Regina for everything they have done in my life and supporting me in every possible way to make my dreams come true. I also want to thank my brother Lushanthan and sister Lushanthini for providing the encouragement and support virtually whenever I felt alone. Huge thanks to my husband Krishanthan for caring me during this challenging phase of life.

Publication

- Linguistic Changes across Different User Roles in Online Learning Environment. What do they tell us?
  - Authors: Lavendini Sivaneasharajah, Katrina Falkner, Thushari Atapattu
  - Conference : Educational Data Mining 2020 (EDM)

- Understanding Students' Learning through User Role and Linguistic Expressions in Online Learning Environment *(Doctoral Consortium)*
  - Authors: Lavendini Sivaneasharajah
  - Conference : ACM Conference on International Computing Education Research 2020 (ICER)

Presentations

- Understanding students' learning in online learning environment
  - Presenter: Lavendini Sivaneasharajah
  - Competition: Three Minute Thesis (3MT) - 2020
  - Award: The Ron Seidel Three Minute Thesis People's Choice Faculty Prize winner for 2020

- Linguistic changes across different user roles in Online Learning Environment *(Poster Presentation)*
  - Presenter: Lavendini Sivaneasharajah
  - Workshop: Australian Learning Analytics Summer Institute (ALASI) 2019

Other Publications

- What Do Linguistic Expressions Tell Us about Learners' Confusion? A Domain-Independent Analysis in MOOCs
  - Authors: Thushari Atapattu, Katrina Falkner, Menasha Thilakaratne, Lavendini Sivaneasharajah, Rangana Jayashanka
  - Journal Venue: IEEE TLT 2020 (Impact Factor: 2.315)

Contents

List of Figures

List of Tables

Nomenclatures

Abbreviations

MOOCs	Massive Open Online Courses
CSCL	Computer-Supported Collaborative Learning
LIWC	Linguistic Inquiry and Word Count
LDA	Latent Dirichlet Allocation
NLP	Natural Language Processing
RQ	Research Question

Notations

N	The total size of the corpus vocabulary
Pi	The proportion of the sample that belongs to class i for a particular node
E	Entropy
A	All possible values for attribute A
v	Any possible values of attribute A
S_v	The subset of S for whose attribute A has value v
$\|S_v\|$	Number of elements in S_v
$\|S\|$	Number of elements in dataset S

Chapter 1

Introduction

Massive Open Online Courses (MOOCs) are large-scale online learning environments that allow learners with diverse academic and professional backgrounds from all over the world to have the opportunity to access courses offered by different universities with internet access (Dillahunt, Wang, & Teasley, 2014). MOOCs are characterized by unlimited numbers of participants ('Massive'); open accessibility where courses are delivered free of charge or impose only low participation fees ('Open'); location-independent as they are available via the Internet ('Online'); and contain didactic content including instructional lectures ('Courses') (Wulf, Blohm, Leimeister, & Brenner, 2014; Clow, 2013; McAuley, Stewart, Siemens, & Cormier, 2010; Vardi, 2012). MOOCs are considered a means for democratizing education (Dillahunt et al., 2014), as they provide an opportunity for every individual across the world to learn and upskill their career irrespective of their background.

MOOCs have grown in popularity in recent years as they have many benefits, such as providing increased access to education, flexible study options, co-construction of knowledge, provide support from a wide range of co-learners and shed new light on course content due to learners' diverse experiences and cultural backgrounds (Lundberg, Castillo-Merino, & Dahmani, 2008; Ferguson & Sharples, 2014). In recent years, MOOCs have become a significant element in the education industry since they offer numerous benefits to the people who are involved in the education sector such as, reducing the overall cost of education programs by minimising the demands for infrastructure and lecture materials (Gaebel, 2014; Hollands & Tirthali, 2014). Furthermore, they are also

a fast-growing research area, where researchers conduct various investigations to extract useful insights about learners.

According to the statistics, by 2021 over 220 million learners globally had registered for a MOOC, including 40 million learners in 2021 alone (Shah, 2021). Furthermore, more than 950 universities are providing at least one MOOC, which has contributed to the growth of MOOCs to over 19,400. However, studies show that only one in every twenty students who enroll in MOOCs complete their studies successfully (Koller, Ng, Do, & Chen, 2013).

Participation in MOOCs seems complex, with students enrolling for varying purposes and intentions, including browsing course content or discovering whether a course is worth pursuing. Some students try to enroll in a range of similar courses in MOOCs to find which is the best fit for their requirements. Students can also enroll just to enhance their knowledge without focusing on final grades or certificates (Koller et al., 2013; Zheng, Rosson, Shih, & Carroll, 2015). This shows completion is not necessarily the only indicator of learning success. Knowing that students may enroll into courses for other purposes, research studies need to explore how to measure learning success through measures other than completion.

A student's success in online learning can be explored using different indicators. Prior literature has used different metrics and cues to measure student success, such as completion (i.e., when a student completes the entire course from start to finish), certification (i.e., when a student earns a certificate of accomplishment), dropout rates (i.e., a student drops out if they do not complete a course) and course grades (i.e., the overall grade in the course). Researchers have studied relationships between these elements to determine student success. Boroujeni et al. (2017) conducted a study to discover the relationship between learners' behaviours and assignment grades. Similarly, Gašević et al. (2016) used log data for predicting student performance. On the contrary, dropouts and at-risk students were predicted using clickstream data (Aguiar, Chawla, Brockman, Ambrose, & Goodrich, 2014). Similarly, several research studies (Aljohani, Fayoumi, & Hassan, 2019; Okubo, Yamashita, Shimada, & Ogata, 2017; Li, Xie, & Wang, 2016; Sharkey & Sanders, 2014; Jiang, Williams, Schenke, Warschauer, & O'dowd, 2014; Kloft, Stiehler, Zheng, & Pinkwart, 2014; Ramesh, Goldwasser, Huang, Daumé III, & Getoor, 2013; S. Crossley et al., 2015; Dowell et al., 2015; Wen, Yang, & Rose, 2014) used MOOC

data for predicting student success. Nevertheless, further investigations are required to understand students' learning beyond grade predictions. This will contribute to further development and enhancement of MOOCs in the future.

This thesis aims to investigate whether analysing learner roles and linguistic expressions that are expressed in discussion forums can be used to understand students' learning in MOOCs. Understanding students' learning is a vast research area, however, this thesis intends to explore how the contributions in discussion forums differ across different learner groups, (e.g., students who obtain different course grades) and finally analyses whether students' knowledge and opinions expressed during the learning process have an impact on their final course grade. Understanding how learners are interacting in the online learning environment can help practitioners to come up with learning strategies that could provide a better learning environment.

MOOCs contain many types of resources to support students in their learning activities. These elements can be categorised as videos, lecture series, reading materials, quizzes, assignments and discussion forums. The discussion forum draws attention among these elements due to its interactivity, where learners have the opportunity to share and interact with limited restrictions. It is a place where students create posts to reflect their original ideas and where reasoning, critical thinking and knowledge of the topic are exchanged in the process of understanding the course content. It also enables learners to comment and analyse the ideas posted by their peers.

Discourse can be defined as a form of language used in both text and talk (Van Dijk, 1997). According to Anderson (2008), discourse enables the learner to come up with their own reasoning and logical thinking by communicating with their peers. In discussion forums, learners engage in discourse on a certain topic to exchange their thoughts. Analysing such discourse is not merely about studying language use but also involves more specific and broader inclusions like who uses the language, how, why and when, emphasizing the importance of the context (Van Dijk, 1997). Discourse analysis goes beyond this and tries to find out answers to questions such as, how does language use influence interactions and how do beliefs control language use? Furthermore, discourse analysis also involves explaining the relationship between language use, interactions and beliefs.

MOOC discussion forums provide a platform for discourse between learners who have diverse background knowledge. The existing literature (Huang, Dasgupta, Ghosh, Manning, & Sanders, 2014; D. Onah, Sinclair, Boyatt, & Foss, 2014; J.-S. Wong, Pursel, Divinsky, & Jansen, 2015) illustrates user participation and their engagement patterns in discussion forums are uneven. Some users actively participate in forums, whereas others are either isolated or use the forum for specific activities (e.g., asking about assignments). These diverse behaviours result in different 'user roles' in discussion forums.

According to the prevailing literature (Strijbos & De Laat, 2010; M. K. Kim, Wang, & Ketenci, 2020; Dowell & Poquet, 2021), several types of user roles have been identified such as 'Pillar', 'Captain', 'Lurker', 'Leader' and 'Over-rider'. These studies emphasise the importance of peer interaction, which is required to identify these user roles. Conversely, role identification can also be built upon the work of post classification methodologies. Several studies (S. N. Kim, Wang, & Baldwin, 2010; Bhatia, Biyani, & Mitra, 2012; Arguello & Shaffer, 2015; Liu, Kidziński, & Dillenbourg, 2016) have been conducted to classify forum posts into different classes (e.g., questions, answers, clarifications, and issues). Hecking et al. (2016) have empirically discovered three major post categories namely: information-seeking, information-giving and other.

Since several roles are identified in the literature, it demonstrates that identifying a set of roles can be based purely on the study data and intentions of the study. This shows that there are no universal roles that can be accommodated by any given research objective. To identify a set of roles that best represent the study data, this thesis analysed the research data. This thesis adopted a grounded theory approach (Vollstedt & Rezat, 2019) to identify emerging roles in discussion forums, which resulted in the emergence of roles including 'information-giver', 'information-seeker' and 'other', similar to the previous study (Hecking, Chounta, & Hoppe, 2017). Hecking et al. (2017) also identified these as key roles.

To identify these learner roles in discussion forum posts, the study conducted a comprehensive literature review on post classification in MOOCs and other domains. Prior literature (S. N. Kim et al., 2010; Bhatia et al., 2012; Arguello & Shaffer, 2015; Liu et al., 2016; Hecking et al., 2016) on post classification shows that several types of features can be extracted from forum posts, such as community-related features (e.g., votes and

views) and linguistic features (e.g., personal pronouns, and analytic skills). The intention of this thesis is to identify the learner role expressed in a discussion forum post in a real-time learning environment. In a real-time environment, it is not realistic to wait for the community-related features to classify students into different roles, as these features can be generated throughout the course and they may change with time. On the other hand, the unstructured nature of interactions that often characterises MOOC discussion forums and the low peer interactions are few reasons for not relying on structural features such as social network positioning. Moreover, the clear focus on linguistic analyses of forum posts highlights what knowledge about learners can be gained from discussion forums alone in addition to the analysis of other modalities such as activity logs.

There is a high chance of course attrition, confusion, and misconception when a student lacks understanding of course content. Real-time role identification is vital as it enables timely identification of information-seekers and the accuracy of information-givers' responses. The work of this thesis is to find more efficient ways to identify roles that could potentially be automated. Therefore, this thesis intends to identify learner roles in discussion forum posts solely based on discourse analysis.

The next focus this thesis strives to analyse is those linguistic expressions that are expressed in learner posts. Linguistic expressions extracted from discussion forum posts have the potential to reveal students' learning. Linguistic cues in forum posts should be studied comprehensively as they can provide further useful insights about students' learning. Several research studies (Sherblom, 1990; Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, & Potts, 2013; Scissors, Gill, Geraghty, & Gergle, 2009; Danescu-Niculescu-Mizil, Gamon, & Dumais, 2011; Postmes, Spears, & Lea, 2000; Huffaker, Jorgensen, Iacobelli, Tepper, & Cassell, 2006; Nguyen & Rose, 2011) have focused on linguistic analysis in online communities. Similarly, studies have also used linguistic analysis in MOOCs to attain various research objectives, such as learner grade predictions (S. Crossley et al., 2015; Dowell et al., 2015; Wen, Yang, & Rose, 2014; S. Crossley, Paquette, Dascalu, McNamara, & Baker, 2016), topic identification (Ramesh, Goldwasser, Huang, Daumé III, & Getoor, 2014; Jagarlamudi, Daumé III, & Udupa, 2012; Xu & Yang, 2015; D. F. Onah & Pang, 2021; Setiawan, Budiharto, Kartowisastro, & Prabowo, 2020), identifying cognitive engagement (Wen, Yang, & Rosé, 2014; C. Beukeboom, 2014; Wang, Yang, Wen, Koedinger, & Rosé, 2015) and investigating learner motivations (Wen, Yang,

& Rosé, 2014). Yet, there remains potential to investigate these linguistic expressions to understand students' learning.

Having a proper mechanism to analyse students' learning accurately makes it possible to provide various benefits for both course providers and learners. Moreover, with the massive amounts of various education data, there is great potential to understand student behaviours, which provides many interesting and valuable research opportunities in this area. This thesis focuses on investigating student learning in MOOCs through user roles and their linguistic expressions.

1.1 Research Questions

Although online learning is not new, the importance of understanding student learning that occurs in non-face to face environments has increasingly gained attention due to the growth of MOOCs in recent years. Moreover, the very recent global COVID-19 pandemic has also forcefully moved face-to-face learning to online environments, which clearly shows the importance of online learning mediums now and in the future.

Given there are several interactive educational components in online learning environments, one of the primary methods to capture students' knowledge and thoughts is by analysing the text messages that are expressed in discussion forums where learners have opportunities to demonstrate their understanding in their own language. Therefore, this research was conducted on the 'discussion forum data' retrieved from AdelaideX courses: 'Introduction to Project Management' and 'Risk Management for Projects'. It is expected that investigating student learning in MOOCs is possible through learner roles and linguistic expressions that can be extracted from discussion forums. To achieve the aforementioned goal, the research conducted three different studies, as presented below:

1. Role Modelling
2. Topic Modelling
3. Linguistic Analysis

1.1.1 Study 1- Role Modelling

The first study in this thesis identifies the learner roles exhibited in the learner posts. The prior literature shows several studies have been conducted to identify user roles from user-generated text; categorising user roles such as Captain, Pillar, Free-rider and Over rider. However, these categorisations require a certain number of peer interactions to be valid. The intention of this study is to identify a learner role in less structured MOOC forums where sufficient interactions among peers cannot be observed. Therefore, this study identified learner roles as 'Information-Giver', 'Information-Seeker' and 'Other' using linguistic expressions alone. This research is guided by the following research questions (RQ):

- RQ1: To what degree of granularity can a machine learning model predict learner roles in discussion forum posts using linguistic features alone?
- RQ2: What are the linguistic features that contribute significantly towards identifying a learner role that is demonstrated in a forum post?
- RQ3: To what extent can machine learning models that rely on linguistic features be used across courses from similar domains?

1.1.2 Study 2 - Topic Modelling

Given the learner roles, the second study of this thesis aims to understand how different learner clusters contribute to different learning topics identified in the learning context. Examining learner clusters is really important, identifying roles such as information-givers, information-seekers, High Distinction information-givers and Fail-grade information-givers. Though topic modelling has been applied in previous research, identifying topics contributed by different learner roles and different learner clusters is still lacking. This study assumes that conducting such analysis has the potential to investigate students' learning as it helps to identify types of learner engagement in different topics: for example, identifying learner topics that are continuously questioned by a learner group. Moreover, comprehensive analysis on learner topics from different perspectives is important to benefit both learners and instructors, as it can contribute

to a better pedagogical approach. To achieve these aims, this study investigates the following research questions:

- RQ1: What are the main discussion topics discussed in learner posts?
- RQ2: What are the main discussion topics discussed in different learner clusters?

1.1.3 Study 3 - Linguistic Analysis

The final goal of this thesis is to propose a set of rules to the learning community (e.g., researchers and instructors) to identify the relationship between linguistic expressions and different learner groups. To do so, initially, this study focuses on predicting learner grades using the linguistic expressions that are extracted from discussion forum posts. Subsequently, the study presents a set of rules from a machine learning model to understand the correlation between the learner grades and linguistic cues. Furthermore, this study also contributes towards defining a set of linguistic characteristics, considering both optional and mandatory participation. This study contributes towards answering the research questions below:

- RQ1: How do linguistic features extracted from students' discussion forum posts contribute to learner grade predictions?
- RQ2: What are the significant rules that can be developed using the linguistic features extracted from discussion forums to identify the likelihood of different learner grades?
- RQ3: What are the significant linguistic features that can contribute to developing linguistic profiles of learners?

1.2 Major Contributions to the Discipline

This section provides a summary of the contributions made by this thesis. The major contribution of this thesis is to improve understanding of student learning in MOOCs through identification of learner roles and linguistic expressions. The MOOC platform contains numerous data about students which are generated by them knowingly and

unknowingly, such as log file data (e.g., time spent, number of problems seen, speed of response), clickstream data and discussion forum posts. Such data provide a detailed overview of students' interactions within the course, but it is questionable whether these data can be used to understand student learning effectively. Investigating how to harness structured and unstructured data can potentially advance our understanding of student learning in online environments. Therefore, the major contribution of this thesis is to determine whether identifying 'learner roles' and 'linguistic expressions' can help to understand student learning in online learning environments, especially in MOOCs.

This thesis uses student roles and their linguistic expressions as a means to investigate students' learning. These factors have been studied in the literature with limited scope and certain restrictions. Several types of user roles (e.g., Follower, Pillar and Captain) prevail in the existing literature; however, there are no definite rules that limit these categorisations. It is important to understand role categorisation highly depends on the given learning environment, as collaboration or communication levels can influence such categories. Lack of strong interconnections and peer communications can result in imprecise learning models during the process of identifying certain roles. Moreover, a lack of collaborative learning among peers has also been reflected in the research data. To address these issues, this thesis implements a predictive model that classifies student roles as 'information-giver', 'information-seeker' and 'other' using the linguistic features from their learner-generated content alone. Generalisability of user roles that were identified using the features (e.g., InDegree, OutDegree and Betweenness) derived from peer interaction is limited, as it's hard to observe enough collaboration among peers in many MOOCs due to its nature. However, a high level of generalisability of learner roles can be achieved when a learner role is identified based only on the language aspects of the given learner post as they do not consider the amount of communication that happens among peers.

Another aim of this thesis is to discover novel and enriched features that can be used in the aforementioned predictive model. Feature extraction and selection techniques are applied to discover those features that contribute significantly towards predicting a student's role in discussion forums. Machine learning techniques will give the best outcome when presented with a carefully engineered feature set. Hence, this study proposes a machine learning model to derive sophisticated features that are not taken into consideration in the existing literature on role prediction.

Several research studies have analysed discussion forum data using linguistic analysis with various intentions, such as predicting cognitive engagement, grade predictions and learner motivation. Nevertheless, research on linguistic expression in MOOC discussion forums to understand students' learning is still in its infancy. Therefore, the thesis will also focus on language aspects to further understand students' learning.

A summary of the major contributions of this thesis is given below:

- *An annotated data set for the learning analytic community* - This thesis contributes an annotated data set, which includes annotated discussion forum posts as information-giver, information-seeker and other. This data set can be used by researchers in different ways, such as to study learner behaviours in discussion forum posts by these learner roles, or to train a predictive model using this annotated data as a training dataset to predict roles with other MOOC data or, as a testing dataset. Moreover, this thesis also presents a detailed description of the annotation process, which can be followed by other researchers to find the set of user roles that lies in their research data.

- *A predictive model for role prediction in discussion forum posts* - A machine learning model has been built to predict learner roles using only those linguistic features that are extracted from leaner posts. Furthermore, this predictive model has been validated in another course to ensure its generalisability to predict learner roles in similar courses.

- *A methodology to identify the significant topics* - This thesis proposes a detailed research methodology to identify discussion topics using an extended training corpus and contributes towards a trained topic model that can be used to identify discussion topics in similar courses. This thesis also describes how these topic models can be used in subsequent semesters to identify discussion topics, ensuring the re-usability of the model.

- *A predictive model based on the linguistic features to predict learner grades.* - A series of predictive models have been built to predict learner grades using only their discourse in discussion forum posts.

- *The proposal of a set of rules based on linguistic features to identify the likelihood of different learner grades.* - A set of rule sets were extracted from a machine learning

algorithm to propose linguistic behaviours of different learner grades. These rules sets are built upon by considering the optional participation of learners. Therefore, these rule sets can be adopted even if a learner does not contribute to discussion forums for some weeks.

- *The proposal of Linguistic Profiles.* - Linguistic Profiles for pass and fail grade learners are presented, based on the learners' discourse. These linguistic profiles are built upon the significant linguistic behaviours that were extracted from the forum posts. Instructors can use the linguistic profiles as a template to intervene, as they can identify the likelihood of learners' final course grades using these linguistic characteristics.

1.3 Thesis in brief

- *Chapter 2 (Background):* The background chapter reviews the existing body of knowledge of roles and linguistics that are identified in online learning environments. It outlines the research on the types of roles that are identified in a collaborative learning environment, followed by post classification methods that were conducted in different forums. This chapter also outlines the potential of linguistic analyses that were conducted in online environments. The later part of this chapter presents social and psychological aspects of the linguistic features used in this thesis.

- *Chapter 3 (Research Design):* Research Design describes the research context, research methodology and provides an overall summary of the studies involved in this thesis. This chapter also outlines the machine learning framework and evaluation methods used in this thesis.

- *Chapter 4 (Research Data):* Research data presents the comprehensive information on the data used in this thesis. More specifically, this chapter outlines the reasons behind the role categorisations and describes the annotation processes that are used to produce the dataset.

- *Chapter 5 (Role Modelling):* The chapter on Role Modelling presents the roles identified in the discussion forums. This chapter implements a machine learning model to predict learner roles using linguistic expressions alone that have been

extracted from learners' discussion forum posts. This section also includes a detailed description of feature extraction, and feature engineering, along with model parameter tuning. This chapter also presents the evaluation results to ensure the validity of the model by applying the model to another learner course.

- *Chapter 6 (Topic Modelling):* This chapter presents the detailed procedures used to identify topics of learner posts. Furthermore, it provides a comprehensive analysis of the relationships between different learner clusters (e.g., information seekers, information givers, pass-grade information seekers, and fail-grade information-seekers) and topics.

- *Chapter 7 (Linguistic Analysis):* This chapter presents the relationship between learner grades and linguistic expressions that are expressed in learner posts. The chapter presents a detailed methodology to extract a set of rules from the predictive model that predicts learner grades. This chapter also presents an analysis of these rule sets and identifies significant linguistic behaviours by different learner grades, considering optional participation. The later part of this chapter visualises the linguistic features against course duration and presents 'Linguistic Profiles' for different learner grades.

- *Chapter 8 (Conclusion):* The conclusion provides a summary of the findings for all the research questions that are identified in this thesis. This chapter also outlines the practical implications of the research findings, and identifies potential threats to the validity of the work. The latter section of this chapter describes potential future directions for the studies presented in this thesis.

Chapter 2

Background

2.1 Introduction

With dramatically increasing popularity, Massive Open Online Courses (MOOCs) have drawn the interest of millions of learners. A MOOC is a large-scale education platform, specifically crafted for distance education, where learners with diverse educational and geographical backgrounds can learn online, typically free of charge. MOOCs have provided a convenient education environment for over 220 million students globally, with over 40 million new learners joined in 2021 alone (Shah, 2021). MOOCs provide various educational settings, notably where technological and instructional artefacts are integrated into the online learning environment to facilitate learning.

Though learning is the first priority of any learning medium, MOOCs have dramatically changed the way the world learns by transforming traditional face-to-face learning into an online learning environment where thousands of global learners can navigate the available resources autonomously (Johnson, Becker, Estrada, & Freeman, 2014). MOOCs also overcome the geographical and temporal limitations that prevail in face-to-face learning.

MOOCs have attracted the attention of education institutions, such that 950 universities around the globe now offer nearly 19,400 courses (Shah, 2021). Figure 2.1 illustrates the growth of MOOCs from 2012 to 2021.

The intention of a course delivered through any learning medium is to enhance the knowledge of learners in some way. However, understanding how students learn in MOOCs is a

vast research area that needs to be studied from different perspectives such as students' contributions on different course topic or analysing their language use. Being high in popularity in the education sector, MOOCs have recorded low retention and/or completion rates (Perna et al., 2014; K. Jordan, 2014), inviting research into the possible ways to investigate students' learning process.

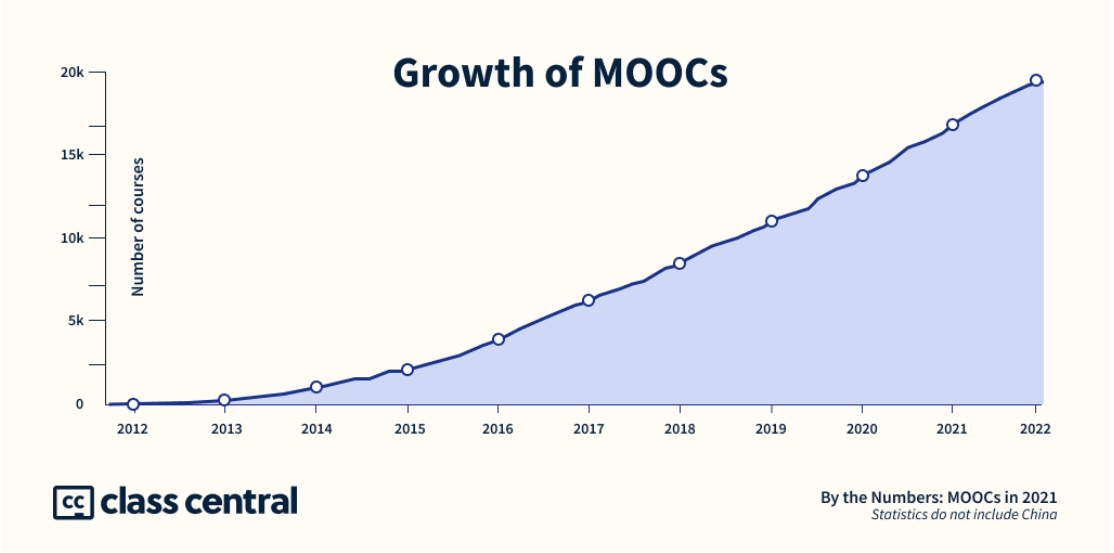


FIGURE 2.1: Growth of MOOCs from 2012-2021 (extracted from (Shah, 2021))

Jordan (2014) conducted an experiment on a series of MOOC offerings and reports that an average of 43,000 students enroll into a MOOC course with a 6.5% completion rate. Similarly, Perna et al. (2014) report only 2% - 13% of students who register in a course attempt the final quiz. It is also important to consider that diversity of measures used in several studies to assess completion rates, such as students who access the last lecture, attempt the final quiz and those who receive a certificate. Furthermore, Ho et al. (2014) report that the completion rates (i.e., course certification) in MOOCs are misrepresented.

Participation in MOOCs seems complex, with students enrolling for varying purposes. For example, students can still access the course materials and participate in the discussion forums without aiming at course completion. Therefore, completion is not necessarily the only indicator of learning success. Knowing that students may enroll in courses for other purposes, it is necessary to explore other perspectives of learning success, beyond completion.

An online learning environment contains many types of resources to support learners in their learning activities. These elements can be categorised as videos, lecture series,

reading materials, quizzes, assignments, learning activities and discussion forums. Students leave footprints in these elements, knowingly or unknowingly, making it possible to investigate the learners' behaviours. A student's footprint can be seen in many MOOC components, such as the time duration spent on learning resources (e.g., videos), interactions on discussion forums and log reports (e.g., views or clicks). These footprints can be investigated to understand the students' learning in MOOCs. Discussion forums have been emerged as a vital element in online learning, as they provide an interactive space for learners to discuss course content. Discussion forum is a place where students create posts to reflect their original ideas and where their thoughts, knowledge of topics, reasoning and critical thinking are exchanged with their peers to understand the course content. They also allow students to comment on and analyse the ideas that have been posted by other students. These learner interactions can happen at various levels, such as learner to learner or learners to instructors, to promote knowledge sharing. According to Anderson (2008), discourse enables the learner to develop their own reasoning and logical thinking by communicating with others. Thus, investigating discussion forums will help researchers and instructors to understand where students are at in the learning lifecycle.

MOOC discussion forums facilitate information exchange between learners who have various backgrounds of knowledge. The existing literature (Huang et al., 2014; D. Onah et al., 2014; J.-S. Wong et al., 2015) illustrates user participation and engagement patterns in discussion forums are uneven. In other words, active learners participate strongly in the forums, whereas non-active learners are either isolated or use the forum for specific activities (e.g., looking for quick solutions to their specific questions, clarification about assignments and answers for assignments). Active and non-active users in discussion forums are generally measured through quantitative measures (Huang et al., 2014; J.-S. Wong et al., 2015). Most commonly, the level of user participation (e.g., the average number of contributions per week or number of posts) is used as a measure to quantify the active and non-active users. Conversely, the contribution they make in forums (i.e., qualitative measures) can also deviate from one to another: in other words, some students can ask questions (i.e., be information- seekers) and a few of them can lead discussion forums (i.e., captains). These diverse user behaviours can lead to different user roles to emerge in discussion forums. However, it is important to understand there is no universal role categorisation as learner behaviours are different from one MOOC to another. To

identify the suitable roles that best describe learner behaviours in a given MOOC learning environment, analysing the research data (i.e., forum posts) is vital. This emphasises the importance of the post classification that has been conducted in the prior literature to analyse author intentions. Post classification can be defined as analysing the forum posts to classify them according to author's intention (e.g., identifying urgent posts or confused posts or leadership posts).

Students' language is another important cue that can be extracted from forum posts. Studies show language can reflect psychological aspects (J. W. Pennebaker, Mehl, & Niederhoffer, 2003; Tausczik & Pennebaker, 2010). Analysing such linguistic expressions can contribute to understanding learners' thoughts and knowledge. For example, it is possible to understand which elements of the course content that presented issues with the help of the learners' language. Furthermore, analysing the effect of language in final grades is also helpful to investigate success and at-risk students.

This thesis assumes understanding students' learning in MOOCs is made possible by inspecting these 'learner roles' and 'linguistic expressions' that are exhibited in discussion forums. A comprehensive background on learner role identification, along with post classification (see Section 2.2.3), is required to understand the existing learner roles that prevail in the literature. It is also believed that investigating the linguistic expressions that are presented in students' forum posts will further contribute to understand their learning, such as identifying language use of at-risk students compared with the language of success students and the way different learner groups (e.g., pass and fail) contributes to different course topics. Therefore, the present work builds on research in: (i) user role identification and post classification in MOOC discussion forums (ii) linguistic analysis of MOOC discussion forums. The following sections discuss each of the topics, illustrating the current status of the field and the research studies undertaken so far on MOOCs.

This chapter is structured as follows: Section 2.2 introduces collaborative learning and describes the roles identified in Computer-Supported Collaborative Learning (CSCL) environments. This section also discusses the importance of post classification, summarising the existing post identification techniques and features used in building predictive models. This section concludes with the motivation to identify the learner roles used in this thesis. Section 2.3 summarises the linguistic analysis conducted in online communities,

followed by the potential of linguistic analyses conducted in MOOCs. Section 2.4 describes the psychological aspects of the linguistic features used in this thesis and Section 2.5 presents the summary of the chapter.

2.2 Role Modelling

This thesis focuses on analysing discussion forums, as they are one of the primary place in MOOCs where learners exchange their thoughts and knowledge of topics. Discussion forums can promote collaborative learning by providing a space for working together in groups or to build a course community to discuss learning content. Such collaborative environments enable the students to raise questions and issues on an unclear course component and also enable the students to support their peers by providing solutions and clarifications. Therefore, it is important to understand how students use collaborative learning environments in MOOCs and the types of roles that have been identified in the existing literature.

2.2.1 Computer-Supported Collaborative Learning

Computer-Supported Collaborative Learning (CSCL) is an established discipline in education that builds upon the concept of collaborative learning and technology support to assist collaborations. It shifts the process of learning from a typical teacher-centred learning environment to a student-centred approach and highlights the importance of collaboration in the learning process. CSCL has been implemented in different levels of education from higher education to middle school students who use computer software to assist their learning (Stahl, 2004). Beyond the education domain, CSCL has been investigated by researchers from various disciplines such as sociology, anthropology, linguistics and communication sciences, who study the language and culture of a community (Koschmann, 1996).

Collaboration has been identified as one of the four critical skills in the 21^{st} century and is considered to be an essential skill for student success as it enhances cognitive development (Webb, Nemer, Chizhik, & Sugrue, 1998; Zhang, 1998). Collaborative learning can be defined as students working together by sharing their knowledge and skills to perform an activity, particularly in MOOCs, to solve a problem or complete a

task (Evans, 2020). Collaboration can play a key role in students' learning in a well-established CSCL, as it can deliver a suitable environment for learners to interact. During collaborative learning, learners can obtain a greater level of conceptual understanding, as they explicitly process the problem with their peers (Darling-Hammond, Austin, Lit, & Nasir, 2003). Moreover, collaborative learning benefits the learning community by encouraging learners' accountability (i.e., responsibility for their own learning), reflective skills, and helps students to actively participate in the learning process, developing their critical thinking ability and improving performance, encouraging students to articulate their thinking, ask questions, increasing their ability to share ideas and solutions, justify their responses and work together in problem-solving (Webb, Troper, & Fall, 1995; Soller, 2001; Roberts, 2005; Baghaei, Mitrovic, & Irwin, 2007). Also, social interactions that occur in a collaborative environment can provide additional ideas to learners that can shed new light on a learning concept (OECD, 2013). These benefits, however, are only achieved by active and well-functioning learning teams (Jarboe, 1996), demonstrating the importance of the learning context.

CSCL is positively associated with many pedagogical approaches such as Distance Education, Discussion, Knowledge Building, Scaffolding and Problem-based Learning (Jeong, Hmelo-Silver, & Jo, 2019). Many tools are available to foster collaborative learning in CSCL. One such tool is discussion forums that are typically available in MOOCs. These discussion forums are available in popular platforms like Coursera[1], edX[2] and Udacity[3].

CSCL is a wide research area that can be studied with different intentions. The interactions and collaborations that occur in a CSCL environment can be examined to study a broad range of learners' skills, such as cognitive and social skills. Knowledge construction (Veldhuis-Diermanse, 2002) and group communication analysis (Dowell & Poquet, 2021) are examples of research foci that have been conducted on a CSCL environment.

In a collaborative learning environment, it is believed that learning is a natural process that happens due to socialisation. In such environments, learners can present, defend and exchange their ideas and beliefs along with questioning others' opinions and knowledge (Srinivas, 2011). This diverse behaviour and different levels of prior knowledge will result in different learner roles, based on the way they behave in the forums. However, it is

[1]https://www.coursera.org/
[2]https://www.edx.org/
[3]https://www.udacity.com/

important to understand not all the CSCL environments are identical, emphasising the fact that different learning contexts need to be examined from different perspectives, such as analysing group formation or language cues to understand the learning process.

2.2.2 Roles in Computer-Supported Collaborative Learning

'Role theory' conceptualises an individual's social behaviour: the way they interact in a group setting is determined by a person's social identities and circumstances (Biddle, 1986). Roles have been studied in several disciplines such as Social Sciences, Education, Psychology and Anthropology. It has become a promising concept in education, especially in CSCL in recent years. It is a paradigm to analyse and understand collaborative learning that happens in a group setting. Apart from understanding the behavioural characteristics of learners results from their social positioning (Biddle, 2013; Burt, 1982), roles on the other hand, can act as prescribed actions (Winship & Mandel, 1983). Roles can be defined as stated functions or responsibilities that guide individual behaviour (Hare, 1994), or an individual's behavioural reaction to a situation caused by communication with peers (Volet, Vauras, Salo, & Khosa, 2017). Generally, roles in a group setting can promote different skill sets, such as group cohesion and responsibility (Mudrack & Farrell, 1995).

According to the literature (Strijbos & Weinberger, 2010), roles in CSCL are looked from two different perspectives, namely scripted roles and emergent roles. Scripted roles are assigned to an individual beforehand to facilitate and shape the collaboration during the learning process (Kollar, Fischer, & Hesse, 2006; De Wever, Van Keer, Schellens, & Valcke, 2010; Weinberger, Stegmann, & Fischer, 2010). In a scripted role environment, each student is apportioned a responsibility or a role and then acts according to the given responsibility (Salazar, 1996). On the other hand, without any prior role assignments, roles that develop spontaneously due to an individual's behaviours and interactions with peers to assist collaborative learning within a group are known as emergent roles. During this collaborative learning setting, an individual's behaviour and interactions develop naturally, caused by their interactions with their peers.

Strijbos and De Laat (2010) present a conceptual framework on roles in CSCL. Their framework comprises eight different roles in small and large group settings. They define four roles, namely Pillar, Generator, Hanger-on and Lurker, in large group settings, with

four different roles, including Captain, Over-rider, Free-rider and Ghost, in small group settings. The study compares the roles in small and large group settings and presents their characteristics in a collaborative environment.

According to the study (Strijbos & De Laat, 2010), 'Captain' and 'Pillar' act as mirror images, where both the roles reflect high levels of engagement and social responsibility in a group interaction. Apart from accomplishing activities they promote the 'togetherness' in collaboration. Another comparison of roles stated in the study is 'Free-rider' and 'Hanger-on', as they replicate similar characteristics: they need group support to complete their individual goals. Contribution to a group activity needs to be stimulated explicitly for a 'Free-rider'. Similarly, a 'Hanger-on' has a great interest in group activity, however with limited contributions that are explained using their own experience. These types of learners tend to achieve high performance in group tasks with very small contributions.

Similarly, the study (Strijbos & De Laat, 2010) compares the 'Lurker' and 'Ghost' in large and small groups, respectively. Their characteristics indicate that they are not involved in the group collaboration. Comparison between the 'Generator' in a large group with the 'Over-rider' in a small group confirms that these two roles make a significant effort to convert the group in the direction of their personal goals, or they try to play an important part in the group goals. They initiate collaboration and continuously emphasise consideration of their previously proposed ideas in accomplishing the group goal.

On the other hand, another set of roles have been proposed by Dowell and Poquet (2021): 'Lurkers', 'Followers', 'Socially Detached', 'Influential Actors' and 'Hyper Posters'. They identify these roles using group communication analysis measures with social network analysis. According to their analysis, 'Followers' are defined as having somewhat above average participation with meaningful contributions, making social impact on the collaboration. Similarly, above average measures across participation levels, social impact and overall responsiveness can be observed in 'Influential Actors' with high internal cohesion, demonstrating self-regulation in learning. 'Hyper Posters' are characterised with high measures on entire group communication analysis measures (e.g., participation, social impact, and overall responsivity). 'Socially Detached' learners are described as having high

participation with high contributions in new information; however, they are ineffective in making their peers respond, indicating less social impact.

In this line of research, another set of roles were introduced by Hecking et al. (2017). The study identifies roles as 'core users', 'peripheral information-givers' and 'peripheral information-seekers'. Apart from the roles identified in CSCL, a set of roles such as 'Resource investigator', 'Coordinator' 'Teamworker', 'Implementer' is introduced by Belbin (2010) in a managerial team work environment. Likewise, several studies have identified various roles that suit their own contexts.

Studies have used different measures and approaches to identify these roles in CSCL environments. To date, most of the existing literature identifies these roles using Social Network Analysis (SNA). Kim et al. (2020) identify 'leaders' in an online discussion forum using social network analysis with three metrics (InDegree, OutDegree and Betweenness). The study identifies how the characteristics of a leader deviate from their peers in three aspects, namely behavioural, cognitive, and emotional engagement. Results show that leaders possess a high value across these aspects compared with non-leaders. The study by Marcos-García et al. (2015) also identifies roles using social network analysis metrics. They evaluate their proposed approach in a case study to identify two different roles (teacher-facilitator, student-animator).

The study by Dowell and Poquet (2021) uses group communication analysis that captures the semantic properties of discourse along with social network analysis. It uses six group communication analysis measures, namely Participation (level of engagement), Social Impact (impact they create within the group to response), Overall Responsivity (semantic similarity with their peers' contributions), Internal Cohesion (consistency of an individual with their previous contributions), Newness (novelty of the contribution) and Communication Density (density of semantically meaningful information). Moreover, their social network analysis measures the learner's position in the network using network measures such as weighted degrees and weighted clustering coefficients.

On the other hand, the indirect blockmodeling network analytic technique has also been used to identify the social roles of students in collaborative learning environments (Medina et al., 2016). The blockmodeling technique is applied in social network analysis to detect the roles in social networks. It reduces the network to macro blocks by grouping them based on their connection patterns (i.e., similar connections). Similarly, Hecking

et al. (2017) have used a socio-semantic blockmodelling approach that integrates the network and content analytics and identifies socio-semantic roles.

Even though several roles have been identified in a collaborative learning environment by various authors, there are no definite rules that say learner roles are only limited to the aforementioned categories. It is important to note that a role that has been identified in a learning context can be referred to by a different name in another learning environment. For example, Strijbos and De Laat (2010) argue that the roles known as 'Captain' and 'Pillar' that are identified in their study are represented as 'social participants' and 'active learners' in the literature (Bento, Brownstein, Kemery, Zacur, et al., 2005). Moreover, it is apparent that not every role can be matched to every circumstance in a learning environment, as it depends on several other contextual parameters and individual personalities. For example, a learning environment with limited levels of collaboration or with limited communication levels can majorly affect identifying user roles like Pillars and Captains. The study by Medina et al. (2016) addresses, how implementing social network analysis alone will fail to identify the collaborative actions that happen at the individual level. The reason behind such limitation is that social network analysis identifies the similarity in roles based on their structural position in a network (i.e., who shares similar structural patterns classified as a single role). On the other hand, the study (Medina et al., 2016) also addresses how identifying complex roles such as Leaders, Coordinators, Active, Peripheral and Missing roles can be inaccurate when using indirect blockmodeling. Their study results confirms that identifying basic roles (i.e., managers or workers) is possible through indirect blockmodeling techniques, however, they will not identify complex roles effectively. They conclude that indirect blockmodeling is effective when there is distinctive behaviour among the roles. Similarly, Hecking et al. (2017) argue that a lack of strong interconnections between the semantic structure and the structure of the information exchange will result in an imprecise blockmodel.

With the aforementioned analysis, it is evident that defining a set of roles in a CSCL environment and selecting an approach to identify these roles purely depends on the MOOC environment and the level of individual contributions that is provided by learners not only to a topic but also with their peers. If there is a limitation in any of the components (e.g., the MOOC environment, individual contributions or a lack of strong collaboration), it will result in an inaccurate role identification in the given collaborative

environment. Therefore, this thesis intends to identify learner roles in the study data, by following the grounded theory approach (Vollstedt & Rezat, 2019). The grounded theory approach emphasises the importance of open coding prior developing a coding framework. It is important to analyse the research data (i.e., forums posts) to define an appropriate set of roles that best describes the given context. The detailed description of the roles identified in the data and the process followed in identifying the roles are presented in Chapter 4.

Moreover, the intention of this thesis is to identify learner role that expressed in their forum posts. This emphasises the importance of post classifications that have been conducted in the existing literature in MOOCs and other domains. Therefore, the background on role identification in MOOCs also includes the prior work conducted for 'Post classification'.

2.2.3 Post Classification

New ways of thinking about discourse and conversational dialogue advanced with the early research on speech acts by Austin (1962) and Searle (1976). Speech act theory was initially developed research in philosophy and sociolinguistics. By addressing the weaknesses in Austin's taxonomy (Austin, 1962), Searle's taxonomy of speech acts classifies illocutionary acts into five different classes: representatives (or assertives), directives, commissives, expressives, and declarations (Searle, 1976). This classifies an utterance/sentence into a set of classes based on three main aspects (purpose, direction of fit and psychological state) of a sentence rather than solely focusing on the structure. This taxonomy has been widely used in the literature and has proven to be a most successful method in speech act classification.

From this point, several classification mechanisms have been evolved, based on Searle's taxonomy (1976) of speech acts. Qadir and Riloff (2011) successfully implemented a sentence classifier using a Support Vector Machine with a polynomial kernel to classify message board posts into the four original speech act classes defined by Searle (1976).

With time, researchers defined with their own speech acts that are tailored to serve different domains. Cohen et al. (2004) proposed a methodology to automatically classify emails based on the sender's intention. They have defined their speech act classes with

the help of Searle's work (1976) and existing work associated with work flow tracking and email speech acts (Winograd, 1987; Flores & Ludlow, 1980; Weigand, Goldkuhl, & de Moor, 2003).

Kim et al. (2010) have examined web forum posts to classify them into different categories. They have identified 12 post categories such as: Question-Question, Question-Add, Question-Confirmation, Answer-Answer and Answer-Correction. The study follows a different mechanism in categorising forum posts i.e., when classifying a given forum post, the study also considers the earlier post. Apart from this, their research study uses lexical features, structural features, post context features and semantic features in determining the category of a forum post. The results show that Conditional Random Field (CRF) performed above baseline results in classification.

The work by Bhatia et al. (2012) on user messages of web forum data is similar to the existing works in the literature. Their work focuses on implementing a classifier to classify user messages into different set of classes (e.g., Question, Solution, Repeat Question, Further Detail). This study extracts content features, user behaviour, structural features and sentiment features for predictions. Using a Logit Model (logistic regression) classifier Bhatia et al. (2012) achieved 72.02% classification accuracy. The study also examined the relative importance of different feature sets by performing a series of experiments. According to the results, the highest and lowest individual performances are obtained for content and sentiment features respectively.

Post classification has also been conducted on MOOC forum data (Arguello & Shaffer, 2015; Liu et al., 2016; Hecking et al., 2016; Wise, Cui, & Vytasek, 2016). In 2015, Arguello and Shaffer (2015) worked on MOOC discussion forum data to describe the purpose of a post with seven different speech acts (e.g., question, issue, issue resolution). With time, researchers started to integrate several types of features to enhance the accuracy of the classifier. The study (Arguello & Shaffer, 2015) also integrates Linguistic Inquiry Word Count (LIWC) (J. W. Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007) features with Non-LIWC features (e.g., sentiment features, temporal features, and votes). The LIWC features were generated from the LIWC tool. These features have been further categorised into several categories such as affect features, cognitive features and linguistic features. On the other hand, Non-LIWC features are categorised into sentiment features, unigram features and text similarity features. These features were

trained and tested using 20-fold cross validation and statistical significance is obtained through Fischer's randomization test (Smucker, Allan, & Carterette, 2007). Moreover, feature ablation analysis was performed on the data set, and it confirms that unigram features and sequential correlation features have a significant impact in the predicting speech acts of a given forum post.

Liu et al. (2016) built a model to automatically annotate discussion forum posts into a set of categories (e.g., Clarification request, Question, Clarification). Post classification is conducted based on the post classification framework presented in the existing literature (Sridhar, Getoor, & Walker, 2014), along with modifications that suit the MOOC environment. A limited set of features to predict the post categories is one of the major drawbacks of this study as it fails to differentiate a few classes from one another: thus, for example, 'Clarification Request' and 'Negative Feedback' possessed similar values for predictive features.

In this line of research, Hecking et al. (2016) have carried out post classification by generalising the categories that prevail in the existing research studies (Arguello & Shaffer, 2015; Liu et al., 2016). The study generalises the existing categories into three classes (information-seeking, information-giving and other). It uses structural (e.g., position in the thread, number of votes) and content related features (e.g., number of question words, question marks) for classification purposes.

Classifying posts into content and non-content posts is another type of post classification study conducted in MOOC discussion forums. A study (Wise et al., 2016) has been conducted to identify content-related and non-content posts using the linguistic features extracted from discussion forums. However, this can be considered as a general categorisation. This study did not capture the specific categories, such as questions, clarifications and information seeking. This will result in delayed instructor interventions due to the time spent on reading every post.

2.2.4 Sustaining CSCL in MOOCs

To capture the overall role of learners, effective communication among peers is of the utmost importance. Without the right amount of interactions, it is not possible to have a well-established collaborative setting in MOOCs. On the other hand, the existence of

a collaborative tool alone is not adequate to foster collaborative learning, as the levels and types of interactions are influenced by many factors. Learner background is one such factor, where learner interaction purely depends on how a learner reacts in discussion forums, whether they wish to continue in a collaborative event, or they just post their thoughts without any peer interactions. If learners lack a wide range of opinions, it is not possible to observe adequate collaborations among them. Moreover, promoting collaborative learning by having continuous instructor interventions is another factor that is responsible for a proper collaborative learning environment. Additionally, the participation level (i.e., mandatory or optional) in a discussion forum that is required by the course can determine the level of collaboration in MOOCs. Course duration, and any lack of social skills on the part of learners are other factors that can affect the amount of collaboration that happens in MOOCs. Likewise, an effective collaborative environment can be built by satisfying these factors to some extent. Therefore, having a discussion forum is not enough for peer interactions: other aspects of the learning context are equally important to have sufficient collaboration in MOOCs.

With the aforementioned emerging roles during CSCL in both small and large group settings, it remains an open question as to what extent and in what ways roles can be identified in less-structured MOOC discussion forum posts. In other words, in less-structured data, communication is either restricted to inadequate levels of interactions or treats the learner replies to a particular thread as taking one single position (i.e., does not capture the multiple positions of replies) or generates insufficient replies to capture the aforementioned roles. This leads to scanty or inappropriate group formation and can eventually misleads the role identification process. In less-structured MOOC data, there is a kind of collaboration happening. However, there will not be prominent teams, therefore these kinds of specific roles are not present in such MOOC settings. Perhaps, a small specific tendency may emerge, but not the full range of roles.

The study of active and rich collaborative learning environments has been widely explored by numerous researchers. However, a collaborative environment with a lack of interactions among peers needs more research. A less collaborative learning environments can be observed in many MOOCs as they are 'Open' (i.e., free or imposing only a small fee) to all learners. Therefore, it is vital to explore the possible ways to understand the learning process in such environments with the cues that are available in the less-structured MOOC data.

In such learning environments, analysing individual posts and contributions in terms of the content of the post can be beneficial. This is one way to gain insight into students' learning in MOOCs. This type of analysis would try to find answers to questions like: How to identify learner needs to retain them in the learning process in real-time? What are the significant differences can be observed between the students who seek for information and give information? How do learning topics deviate from the students who seek and give information?

2.3 Linguistic Analysis

Language is an important human skill and is considered to be one of the greatest powers of mankind. Language has the ability to distinguish humans from other creatures, as it provides the capability to create words and sentences. Generally, language can be described as a sign system that is used by humans to communicate their thoughts and feelings to another (Shahhoseiny, 2013).

Studies (Shahhoseiny, 2013; Robins, 2014) define linguistics as the *'scientific study of language'*. Linguistics has many related sub fields such as descriptive linguistics (study associate with description and analysis of the way a language operates at a given time), historical linguistics (study about language development in history) and comparative linguistics (study that compares one language with another) (Robins, 2014). According to Chomsky (1986), linguistics aims to describe the content of human language using suitable terms.

Figure 2.2 illustrates the major levels of linguistic structure namely: phonetics, phonology, morphology, syntax, semantics and pragmatics (or discourse) (Mahmood et al., 2016; Mahmood, 2019). Phonetics and phonology focus on linguistic sounds. The morphological level of a language closely twinned with the formation of words. Structural relationships between words are studied using syntax; whereas semantics studies the meaning of words. Pragmatics focuses on the connection between the meaning of the language with the intentions of the speaker. Finally, discourse studies linguistic units larger than a single utterance.

Linguistic analysis is described as scientific study of language, which considers at least one or more of the aforementioned linguistic structures (e.g., syntax and semantics). It

aims to analyse a language and find ways to use the knowledge further in understanding communication, eventually understanding the human mind. Examining such linguistics of a text can help to understand the knowledge of an individual to a greater extent. Therefore, this thesis includes a comprehensive literature review on the linguistic analyses that have been conducted in online communities, especially in MOOCs.

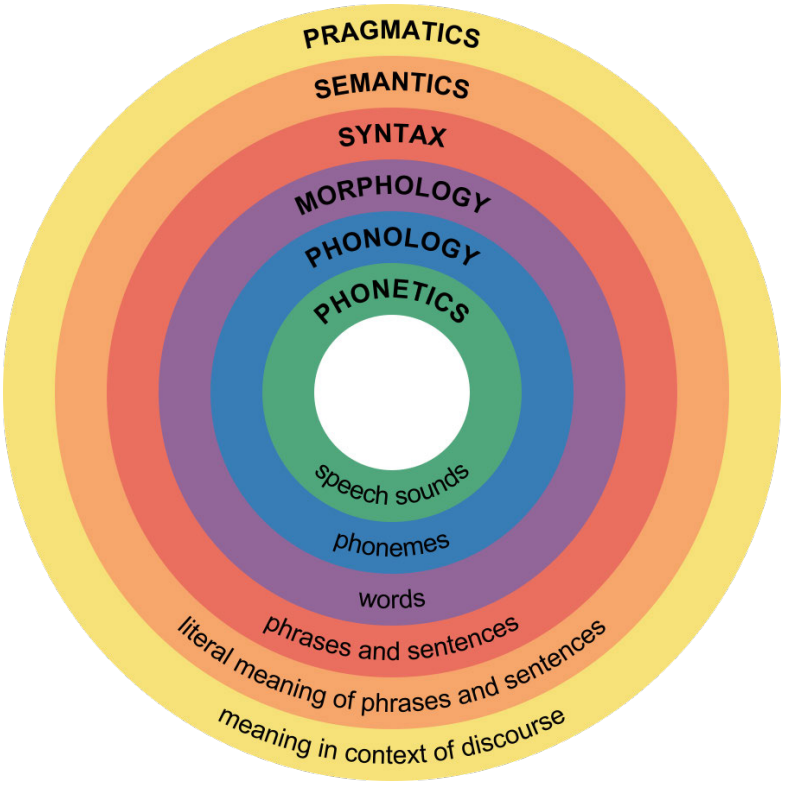


FIGURE 2.2: Major levels of linguistic structure (extracted from[4])

2.3.1 Linguistic Analysis in Online Communities

For decades, researchers in the discipline of linguistics have explored the language changes in many different spheres, from historical linguistics to sociolinguistics. Research scholars believe exploring linguistic expressions and temporal changes in a user's language will provide useful insights to research communities. Linguistic research has taken various paths over time in order to exhibit correlations between the linguistic features and other aspects such as historical change, community norms (Danescu-Niculescu-Mizil et al., 2013; Postmes et al., 2000), and user lifespan (Danescu-Niculescu-Mizil et al., 2013).

[4]https://commons.wikimedia.org/wiki/File:Major$_l$$_e$$_v$$_e$$_l$$_s$$_o$$_f$$_l$$_i$$_n$$_g$$_u$$_i$$_s$$_t$$_i$$_c$$_s$$_t$$_r$$_u$$_c$$_t$$_u$$_r$$_e$.svg

There is extensive research (Scissors et al., 2009; Gonzales, Hancock, & Pennebaker, 2010) in the literature grounded on Communication Accommodation Theory (Giles, Coupland, & Coupland, 1991). This field of research aims to prove that people who are involved in a communication will portray linguistic similarity (i.e., word choice, length of discourse, syntax) to that of their communication partner during discourse. In this line of research, Scissors et al. (2009) investigated the linguistic adaptations that happen in text-based conversations. The results indicated, linguistic similarity in conveying positive emotions, usage of past and future verb tenses, and noun phrase references observed in high trusting pairs. Similarly, Gonzales et al. (2010) have created a linguistic style matching algorithm to predict two aspects of social dynamics, namely 'cohesiveness' and 'task performance' in small groups. The study has identified several relationships between the social dynamics (i.e., group cohesiveness and task performance) and language characteristics (e.g., word count, future-oriented language). 'Word count' demonstrated a positive relationship with the group cohesiveness. Likewise, future-oriented language (e.g., could, will) shows a positive relationship with task performance. Conversely, a negative association is observed between achievement-oriented language (e.g., ability, work) and task performance. The work by Niederhoffer and Pennebaker (2002) shows that linguistic style matching happens at the conversational-level and turn-by-turn level during interactions between two individuals, demonstrating that, unconsciously, changes occur at linguistic-level when an individual is in synchronisation with another during a conversation. In addition to this, work by Danescu-Niculescu-Mizil et al. (2011) develops a probabilistic framework to calculate the style accommodation in Twitter conversations.

Another facet of linguistic research is to identify the correlation between user lifespan and their language use (Danescu-Niculescu-Mizil et al., 2013; Nguyen & Rose, 2011). Nguyen and Rose (2011) have investigated the relationship between community membership and language use. Their findings confirm that long term participants express their opinions using forum-specific jargon and show highly informal linguistic style in their conversations. Patterns of language change with time is analysed using the Kullback-Leibler divergence matrix (Kullback & Leibler, 1951; Kullback, 1997), and Spearman's Rank Correlation Coefficient to calculate the similarity between two-word distributions. Furthermore, the study by Danescu-Niculescu-Mizil et al. (2013) also examined language changes that happen with time by comparing word distributions across consecutive weeks. The results show that incorporating singular first-person pronouns such as

'I', 'myself', 'me' will decline with time as participants write many reviews to the online community. This might be an indicative of the user's increasing involvement with the community. Moreover, the study reveals users use more context-related vocabulary with time.

Similar to this strand of research, Huffaker et al. (2006) investigate similarities in text-based messages across time in an online community. The results show that there is a divergence in the members' language with time. A variety of techniques have been explored to compute the similarity scores, such as: Spearman's Correlation Coefficient, Zipping and Latent Semantic Analysis.

A separate line of work has investigated the relationship between language use and community norms (Danescu-Niculescu-Mizil et al., 2013; Postmes et al., 2000). They believe strongly that becoming a core member and being committed to a community can be measured by a user's conformity to community norms. In other words, the more a user adopts the community norm, the more they become a core member.

One prominent work in this line of research was conducted by Danescu-Niculescu-Mizil et al. (2013). They examined user reactions to community norms in online communities. This study proposes a framework to observe linguistic change in large online communities, both at the individual-level and community-level. The study uses a dataset from two large online review communities (RateBeer and BeerAdvocate)[5]. The study data is suitable to investigate linguistic change since it is from a decade old online community with active participants, where an individual writes over 100 reviews. Their findings confirm that a user follows a lifecycle, where at the beginning of the lifecycle a user adopts the language used in the community; however, at the later part of their lifespan, they are resistant to change in accordance with the community norms.

A further study (Geddes, 1988) reveals that participants' social identities are articulated via communication and their degree of contribution to a work group is demonstrated by their language use. Building on this, Sherblom (1990) investigates workers' involvement with an organisation. The study examines the relationship between the communicative function (e.g., a request for information) of an email message and the use of personal pronouns. He examined these communicative functions using linguistic markers, especially the use of personal pronouns in emails. The results show that use of personal pronouns,

[5]http://snap.stanford.edu/data/

especially the 'I', 'We' and 'You' categories, had a significant impact on categorising communication functions.

2.3.2 Potential of Linguistic Analysis in Massive Open Online Courses

Given the great potential of linguistic analysis in different online communities, research scholars also have attempted linguistic analysis in MOOCs. Since the data from MOOCs have the potential to reveal new insights, different aspects of learning have been studied by several research studies through learners' language, such as identifying learner motivation, cognitive engagement, topic modelling, and language change with time. The richness of the data obtained from MOOCs provides diverse research opportunities. The section below presents a range of linguistic analysis conducted with diverse motivations in MOOC environments.

2.3.2.1 Investigating Learner Motivation and Cognitive Engagement

Detecting cognitive engagement helps to identify the degree of effort involved in investigating, analysing and reasoning about the learning content. With more effort in processing learning materials, more learning is achieved (Chi & Wylie, 2014). Moreover, several studies demonstrate that learner contribution is a vital predictor in knowledge construction (E. G. Cohen, 1994; Barab & Duffy, 2000). For example, studies indicate that learning gain is correlated with the proportion of words in a dialogue (Core, Moore, & Zinn, 2003) and the number of words per utterance (Rosé, Bhembe, Siler, Srivastava, & VanLehn, 2003). Another study (C. J. Beukeboom, Forgas, Vincze, & Laszlo, 2014) shows that the degree of cognitive inferences can be revealed through the different levels of language abstraction. Therefore, it is important to analyse the learner contribution in order to better understand a student's learning in an online environment.

In this line of research, Wen et al. (2014) have built a machine learning model to detect the level of 'learner motivation' using forum posts. The study utilises five different domain-independent motivation cues: apply words (i.e., synonyms for apply, use), need words (i.e., words that demonstrate participants' needs and goals), cognitive words from LIWC, first person pronouns, and positive words, as input features for the model to detect learner motivation.

The study (Wen, Yang, & Rosé, 2014) also conducted experiments to understand 'cognitive engagement', which reveals the determination of a student to interpret, analyse and reason about the lecture content. The study used 'language abstraction' (C. Beukeboom, 2014) as an indicative measure to understand the cognitive engagement of the learner. Each word from learner posts was matched with the words obtained from the abstractness dictionary by Turney et al. (2011) to calculate the average abstraction for each user post.

A similar study (Wang et al., 2015) was conducted to investigate the 'cognitive behaviour' of a learner in discussion forums. The study used three labels, namely Active, Constructive and Interactive, to categorise the discussion forum post. The study built a Logistic Regression classifier using bag-of-words as the feature space, to predict these three different cognitive behaviours and obtained approximately 75% accuracy for each class. Using linear regression models, the study concluded that active and constructive behaviour in discussion forums will result in a significant learning gain.

In addition to cognitive engagement, 'learner engagement' in MOOCs is another type of research path that has been conducted by analysing the MOOC discussion forums. It is believed student engagement plays a major role in student success (Kuh, 2003). Several studies (Ramesh et al., 2013; Kizilcec, Piech, & Schneider, 2013) have conducted research on analysing learner engagement. Kizilcece et al. (2013) implemented 'cluster analysis' to identify different engagement and disengagement patterns in three computer science courses. They named emerging clusters of students as completing, auditing, disengaging, and sampling learners, and provided recommendations depending on the learning trajectories in each cluster.

2.3.2.2 Performance Prediction

Due to low completion rates and to promote retention in MOOCs, a series of research studies have been conducted to understand the learning gain. This core idea has been explored in research studies under different names, such as predicting completion, predicting students at risk, performance prediction and dropout predictions. In addition to this, research studies have also been conducted to identify the factors that correlate with student learning success.

Performance Prediction using Interaction Data

A substantial amount of studies (Palmer, Holt, & Bray, 2008; Cheng, Paré, Collimore, & Joordens, 2011; Okubo et al., 2017) have conducted experiments to predict student performance based on various measures. The studies have investigated the learning process of students and predict their performance based on the existing data that prevail in the learning environment. Such investigations have been performed using different techniques, such as classification and regression (Hämäläinen & Vinni, 2011).

These existing data can be extracted based on a single source (i.e., predicting grades based only on clickstream data) or a combination of various sources (i.e., predicting grades based on clickstream data and existing course marks). Romero et al. (2008) summarise a vast range of data that are available in online education platforms, such as information on the number of times a module has been read, and the number of messages read on the forum. Using these quantitative measures, studies have conducted experiments to predict student performance and determine the factors that can be positively correlated with students' learning.

To begin with, research studies have conducted experiments to identify the correlation between learner participation and learner performance. The work conducted by Patel and Aghayere (2006) defines the participation in forums as the number of times a user posted and read the posts in a discussion forum. Their correlation analysis with course grades indicates students' participation has a positive impact on the course grade in two civil engineering on-campus courses. Similarly, several other studies (Cheng et al., 2011; Palmer et al., 2008) found a positive relationship between the participation and assessment scores using course data offered at the undergraduate-level. Using regression analysis, Cheng et al. (2011) demonstrated that a positive relationship can be observed between the number of posts posted by students and assessment scores. Moreover, apart from the number of new posts posted in the discussion forums, previous academic ability also significantly correlate with final grades (Palmer et al., 2008).

The aforementioned research studies have used quantitative measures (e.g., number of posts, number of reads) to discover the correlations with course performance using various techniques. Cheng et al. (2011) have used regression analysis to examine the relationship

between discussion forum participation and course performance. Using multiple regression analysis, Palmer et al. (2008) analysed how the participation in online discussions can affect student course performance.

Prior work has also applied correlation analysis to determine the factors that impact student success in MOOCs. To understand these factors, studies have used both surveys and behavioural data available in MOOCs. A survey conducted by Gutl et al. (2014) investigates the motivations for MOOC enrolment alongside reasons for dropping out. Several other studies have gone beyond the surveys and used log data from MOOCs to perform correlation analysis to determine the factors that affect learner gains. Studies suggest peer influence (measured using different parameters such as topic modelling, or reply interactions) (Yang, Wen, & Rose, 2014), student behaviour (e.g., thread starter, or post length), social positioning in discussion forums (Yang, Sinha, Adamson, & Rosé, 2013), and sentiments expressed in discussion forums (Wen, Yang, & Rose, 2014) have all been correlated with student success in MOOCs.

These correlation studies help to understand the factors that affect student success; however, building a predictive model is vital in online learning environments, especially in MOOCs, due to the vast amount of enrolments and the difficulty of investigating students' learning compared with more controlled learning environments.

Several studies (Aljohani et al., 2019; Okubo et al., 2017) have used log data for such predictions. The work conducted by Aljohani et al. (2019) predicts students at-risk in a virtual learning environment using the Long Short Term Memory (LSTM) model. Similarly, Okubo et al. (2017) built a Recurrent Neural Network (RNN) along with a LSTM model to predict student grades using the log data. These log data are captured using a series of necessary activities such as submitting reports and slide views. However, applying a similar framework in a MOOC where participation is not mandatory, needs to be validated.

The work by Ashenafi et al. (2015) predicts student performance using the multiple tasks that were apportioned to students during the course. Moreover, it was also possible to predict the at-risk students using in-semester performance data (e.g., quiz grades, or weekly homework grades) (Marbouti, Diefes-Dux, & Madhavan, 2016).

On the other hand, studies have been conducted on MOOCs to understand learning gain. Studies show learner performance can be predicted using the engagement patterns observed in MOOCs. Using a binomial logistic regression model, the study by Bonafini et al. (2017) investigates how a student's achievement is affected by their engagement level on videos and forums. The results depict the likelihood of learner achievement increases with learner engagement in videos and discussion forums. Similarly, the study by Li et al. (2016) proposes a prediction model using prior students' performances. Using regression and back-propagation neural network methods, the experiment was conducted on course data which span across 15 weeks, demonstrating the vast amount of data used for the prediction.

Moreover, Sharkey and Sanders (2014) predicted retention and attrition in MOOCs using the log data such as time spent, and view counts. Using logistic regression, Jiang et al. (2014) predicted the likelihood of students receiving certification for course completion. Similarly, a study by Kloft et al. (2014) used clickstream data to predict the dropouts using a Support Vector Machine (SVM).

Another study predicts student performance using learner engagement in MOOCs (Ramesh et al., 2013). It models the student engagement using Probabilistic Soft Logic along with behavioural features, subjectivity scores, polarity scores and temporal features (e.g., last view and last vote), and then further deployed these features to predict student performance.

Performance Prediction using Linguistic Analysis

Though several studies have used various non-linguistic measures to predict student success in MOOCs, few studies have analysed the content of the discussion forums and used the linguistic features as predictors as a whole or in part. It is demonstrated that Natural Language Processing (NLP) has the potential to contribute towards performance prediction in MOOCs (C. Robinson, Yeomans, Reich, Hulleman, & Gehlbach, 2016). Using this computational approach, the linguistic properties of the text are analysed to understand human language. In this line of research, the work by Crossley et al. (2015) analysed students' discussion forum posts to predict course completion. Their study used linguistic features extracted from linguistic tools, namely the Writing Assessment Tool (McNamara, Crossley, & Roscoe, 2013), the Tool for the Automatic Analysis of Lexical

Sophistication (Kyle & Crossley, 2015), and the Tool for the Automatic Assessment of Sentiment and performed multivariate analysis of variance (MANOVA) analysis to identify the significant differences between completed and non-completed students. High essay scores, a high vocabulary range, high cohesion, and more domain specific words are some significant indicators that were observed in those students who successfully completed the course.

Similarly, a study by Dowell et al. (2015) examined academic performance using discourse analysis. The study used the linguistic tool known as the Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) to analyse student posts in five discourse dimensions. According to their results, abstract language, simple syntactic structures, cohesive integration, and descriptive discourse are observed in the students' posts who performed better.

Furthermore, sentiment analysis (Wen, Yang, & Rose, 2014) has been analysed in MOOCs to predict student success. The study by Wen et al. (2014) observed an association between sentiments expressed in forum posts and dropout rates. However, the study also pinpoints that the results across the courses are not consistent. Therefore, it is important to analyse sentiment analysis carefully while making conclusions, as sentiment words can be used in different ways in different courses. For example, in a Fantasy course, words such as devil, death, horror are associated with the characteristics in the fiction. These negative words actually represent the learner engagement because they are describing the fantasy related information.

Robinson et al. (2016) analyse pre-course open-ended responses revealing students' opinions about perceived usefulness of the course material. The study concludes that an n-gram model performs better than a simple demographics-only model in predicting those students who will complete the online course.

Studies have also combined click-stream data with language features to investigate MOOC completion (S. Crossley et al., 2016). The study by Crossley et al. (2016) conducted a multivariate analysis of variance (MANOVA) to examine the significant differences between the discourse of students who successfully completed the course and those who did not. Their study also performed Discriminant Function Analysis (DFA), a statistical procedure for prediction. Their findings confirm that click-stream data were the strongest

predictors of completion but NLP features, in particular the post length, entities in a forum post, the overall quality of the written post, lexical sophistication, cohesion between posts, and word certainty were also considered as strong predictors of completion.

2.3.2.3 Topic Modelling

Topic modelling is a popular computational approach for investigating text data to best represent the underlying topic of a given document. A topic model can be applied to both structured and unstructured data for various purposes, such as information retrieval, text clustering, text classification and recommendation systems. Topic models are unsupervised approach that does not require manually labeled data. With topic models, underlying hidden topics are discovered across several domains from user generated discourse. Social media (Xie, Zhu, Jiang, Lim, & Wang, 2016) and commercial websites (Rossetti, Stella, & Zanker, 2016; Dupuy, Bach, & Diot, 2017; Westerlund, Mahmood, Leminen, & Rajahonka, 2019) are two examples where topic modelling has been utilised in addition to online education platforms. In an online education platform, the prime goal of a topic model is to identify latent topics that are discussed by students, eventually utilising the outcome to accomplish several tasks such as pedagogical design (D. F. Onah & Pang, 2021) and course recommendation (Apaza, Cervantes, Quispe, & Luna, 2014).

For many decades, course materials have been designed by lecturers and instructors and can be repetitive for subsequent years. However, with high course enrolments and difficulties in one-to-one monitoring, it is important to understand the students' perspectives on the course content. With diverse learning demands, it is the responsibility of the course designers to understand and alter the course content instead of providing a pre-designed static course.

While designing a course, it is vital to note: have these course content been understood by students, whether there are any sections that are less understood, what kind of topics have been discussed often, and content that has been questioned frequently. Answering these questions in a systematic way will help to investigate student learning to a certain extent and eventually accommodate diverse student needs. Moreover, investigating the content of the student discussions will help to unwrap the learners' perspectives, their understandings, and issues. This will provide informative information about student learning. In doing so, the first stage to address this problem is by implementing a topic

model to understand the topics that are discussed in the discussion forum in relation to the course materials. Therefore, this section reviews the literature on topic modelling that have been conducted in MOOCs.

The most frequently observed goal of topic modelling in the literature is to investigate the key themes expressed in texts. In this line of research, the work is conducted on a Cartography MOOC dataset, applying a Topic Modelling Tool, which was designed to use by non-experts to extract key topics in discussions (A. C. Robinson, 2015). Topic modelling has also used to predict student survival (Ramesh et al., 2014). Their study used SeededLDA (Jagarlamudi et al., 2012) in discussion forums to extract the predefined desired topic categories. The study by Xu and Yang (2015) applied Latent Dirichlet Allocation (LDA) based topic modelling to calculate the similarity among learners in terms of course content. Learners' similarity on the course forums was considered as one of the parameters for recommending study partners in MOOCs.

LDA is a state-of-the-art generative probabilistic model introduced by Blei et al. (2003), an extension of Probabilistic Latent Semantic Indexing (PLSI). It overcomes the limitations that exist in other topic models, such as latent semantic indexing (LSI) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) and probabilistic latent semantic indexing (PLSI) (Hofmann, 1999). For example, all meanings of a word are represented by the same vector in LSA whereas multiple meanings of a word explicitly represented in LDA (Rus, Niraula, & Banjade, 2013; Niraula, Banjade, Ştefănescu, & Rus, 2013). LDA is an unsupervised approach to mine semantic information from text data. Using an LDA topic model, a corpus can be represented with a random mixture of latent topics. A topic is a theme that is denoted with a word distribution that often co-occurs in the given corpus. The LDA model identifies a topic for a document according to the co-occurrence of words in the given document. It represents the document as a mixture of different topics with their associated weights. Figure 2.3 presents the graphical representation of the LDA model.

Course design is another application that uses topic modelling, where the outcome of topic models can contribute towards pedagogical design and organising the course content. The work by Onah and Pang (2021) applied the LDA topic model for topic extraction from icebreaker discussions. The insights from the topic model were applied to structure the content of a MOOC on the FutureLearn platform in a logical way.

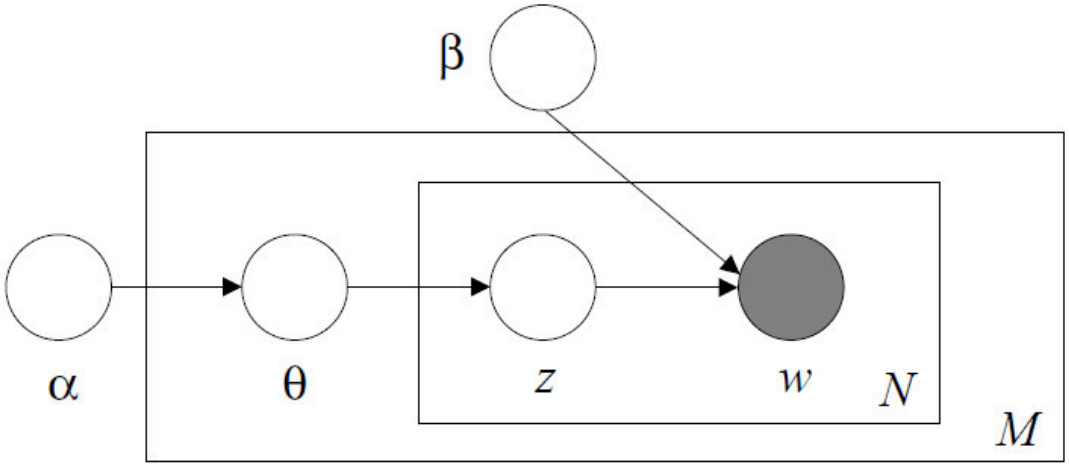


FIGURE 2.3: The graphical model of LDA (extracted from (Blei et al., 2003)). The outer box M represents documents, while the inner box N represents the repeated choice of topics and words within a document.

Probabilistic Latent Semantic Analysis (PLSA) is another topic modelling technique that can be applied to extract the topics from the forum discussions. The study by Setiawan et al. (2020) examined the forums from a Learning Management System (LMS) using the PLSA topic model. Their study also compares the results, with the outcomes of different topic modelling techniques that were obtained from other study (Rajasundari, Subathra, & Kumar, 2017). The results demonstrated that the LDA topic model outperforms other topic models.

Topic modelling has also being applied for recommending courses (Apaza et al., 2014) and resources (W. Kuang, Luo, & Sun, 2011) in learning environments. The study by Apaza et al. (2014) recommends courses based on the grades obtained in the college. Their idea behind using a topic model is to extract topics from both MOOCs and the college course syllabi. Their study used an LDA topic model to investigate the semantic structure of each course and the output of the LDA is used as one of the feature vectors in the recommendation system.

Topic modelling is also used to understand learner engagement. Yan et al. (2019) examined learner engagement in terms of topics, sentiments and content similarity by using content analysis and topic modelling techniques. Their study applied topic modelling to Teach-Out discussion forum data to extract the emerging topics for examining student engagement in Teach-Out. The study also investigated the forum similarity relevant to the Teach-Out content using latent semantic analysis.

Moreover, topic modelling has been utilised to capture the key topics of different clusters. The work by Ezen-Can et al. (2015) applied a k-medoids clustering algorithm to categorise similar discussion forum posts into different clusters and later, applied an LDA-based topic model for topic extraction in each cluster.

Another facet of topic modelling is to connect student discussion forum posts with the corresponding lecture content (A. W. Wong, Wong, & Hindle, 2019; Atapattu & Falkner, 2016). Wong et al. (2019) examined course materials using both unsupervised and supervised variants of LDA on five different Coursera courses related to Computer Science and Software Engineering. Using topic modelling, the study extracted topics from the course materials. To trace back discussion forums with course content, the study deployed the topic models trained on course content to deduce topics from forum discussions where each post is represented by a distribution of topics with its corresponding weights. Similarly, Atapattu et al. (2016) implemented an LDA topic model on two Coursera courses to label the topics extracted from the discussion forums. The study implemented an LDA topic model along with a Naive Bayes classifier on discussion forum data for topic labelling.

Another study (Vytasek, Wise, & Woloshen, 2017) demonstrates that interpretability of the extracted topics improves significantly when applying a topic model to both content and non-content posts. This pre-categorisation improves the understanding on frequent terms like 'quiz' and 'explain'. For example, if the term 'quiz' appears in two different topics, it is easy to differentiate whether it is a conceptual question or how to access the quiz.

Apart from mining topics from forum discussions, LDA has been applied in mining surveys with open-ended questions (Nanda, Hicks, Waller, Goldwasser, & Douglas, 2018; Buenaño-Fernandez, González, Gil, & Luján-Mora, 2020; Nanda, Douglas, Waller, Merzdorf, & Goldwasser, 2021). An LDA topic model is applied to open-ended responses from a post-course survey to understand learners' opinions about participation certificates (Nanda et al., 2018). Similarly, a recent study by Nanda et al. (2021) analysed open-ended learner responses from several MOOCs to understand the important characteristics of MOOCs to improve learner experience. The study used an LDA topic model to extract key topics in the learner responses.

Apart from MOOCs, numerous studies have applied LDA topic models to extract key themes from textual data to identify key topics from educational blogging platforms (X. Kuang, Chae, Hughes, & Natriello, 2021), to understand public discourse about MOOC-related topics on mainstream media (Kovanović, Joksimović, Gašević, Siemens, & Hatala, 2015), and to visualise topics from student interactions (Zarra, Chiheb, Faizi, & El Afia, 2018).

Literature demonstrates several topic models exist, such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990), probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and have been proven to be most successful approaches for text summarising across several domains. However, to date, the LDA topic model is considered to be the most popular model and has been broadly studied across several domains (Calheiros, Moro, & Rita, 2017). Several research studies have proved it to be more accurate than other models. An LDA model can be implemented using several toolkits such as the Machine Learning for Language Toolkit (MALLET)[6], Gensim[7] and the Stanford TM toolbox (TMT)[8].

A research study conducted on a Facebook conversation datasets (Albalawi, Yeap, & Benyoucef, 2020) reveals LDA and Non-Negative Matrix Factorization (NMF) methods deliver more coherent and higher-quality topics than other topic models. It can also be observed that LDA topic models provide better descriptive topics than other topic models. Additionally, LDA provides more meaningful and logical topics than NMF methods and it also outperforms the NMF method and other topic models, demonstrating the LDA topic model is more suitable for constructing a topic model to gain both better performance and more meaningful topics.

The literature demonstrates that LDA is a successful topic modelling approach to extract key themes from user-generated text. Therefore, this thesis applies LDA to train a topic model with lecture content and another semester's discussion forum posts. This trained topic model was applied onto the study data to extract the topics from each forum post (see chapter 6). Moreover, the study is not only limited to topic identification but also analyses the topics expressed in different learner clusters, such as information-seekers and information-givers. The outcome of the analysis can help instructors to

[6]http://mallet.cs.umass.edu/
[7]https://pypi.org/project/gensim/
[8]https://downloads.cs.stanford.edu/nlp/software/tmt/tmt-0.4/

investigate student learning in MOOCs such as exploring the discussion topics that were less understood by learners, and frequently questioned.

2.3.2.4 Linguistic Changes in the Education Context

Linguistic changes in an education context has been studied through a limited set of linguistic markers. The study by Adaji and Olakanm (2019) presents how four dimensions of emotions and sentiments (analytic, authentic, clout and tone) change over five years for Stack Overflow users. Recently, Dowell et al. (2017) conducted a study on MOOC data to identify the conversion in learners' language and discourse characteristics over time. For this purpose, they used a mixed-effects modelling methodology on five different courses extracted from the Coursera platform. The researchers focused on two main discourse communication trends: message relevance and linguistic complexity. Message relevance is computed by analysing the relevance of a message to that of the course video transcript, and linguistic complexity is calculated by Coh-Metrix's (McNamara et al., 2014) Flesch-Kincaid reading level measure (Klare, 1974).

This existing work shows studies related to MOOC discussion forums that investigate temporal changes in a learner's language are still in their infancy. Even though preliminary work on linguistic changes has been conducted in other online communities, a lack of work has been conducted in MOOCs. It is said that a learner's language changes with time. In particular, discourse in discussion forums will be topic-oriented and reflective of deep learning with the subsequent offerings of a course (Dowell et al., 2017). Nevertheless, investigating linguistic change across grade distributions have not been addressed.

To address this gap, this thesis will explore how linguistic expressions change during a course between two different student grades, namely pass and fail. This thesis will also explore the differences in the language between those students who obtained pass grades and those who failed.

2.4 Language and Discourse

This section describes the comprehensive analysis of psychological aspects of linguistics that are extracted from linguistic tools. This thesis uses linguistic features extracted from the Linguistic Inquiry Word Count (LIWC) tool (J. W. Pennebaker, Chung, et al., 2007). Therefore, it is important to understand how these linguistic features are connected to the psychological aspects, which eventually helps to understand learners' language in forum posts.

In understanding a user behaviour, there is an increasing interest in the literature in investigating the language and discourse of a user in many spheres, from psychological to cognitive, to emotions and many others (D'Mello & Graesser, 2012; McNamara et al., 2014). Although several measures can be used to investigate user behaviour such as, log data and social network measures, there remain many important paradigms of users that can be studied only through language and discourse.

Language is used as a tool between two individuals where one individual explains their inner thoughts and feelings in a series of words that can be understood by another. It is believed that the way we think has a high impact on the words we use in daily life. It is important to understand the psychological aspects of words while processing language, as they reveal more information once the psychological aspects are understood, such as an individual's thought process, thinking styles, attentions, intentions and their emotional states (Tausczik & Pennebaker, 2010).

Analysing user language is not an easy process, as it requires a deeper understanding of language cues. Although understanding a learner's language is possible through manual content analysis, it requires a substantial amount of time and patience to investigate the discourse. Moreover, it is not practical to manually analyse the discourse due to the large scale of data, particularly MOOC textual data. To overcome these obstacles, researchers are applying automated linguistic analysis through automated language tools that have been designed after rigorous evaluations. These automated language tools analyse both shallow-level and deeper-level discourse, enabling the researcher to choose the options that best address their research problems. It is proved that not only deeper-level discourse but also shallow-level discourse can provide more insightful measures about a user.

There are multiple automated language analysis tools[9] available in the academic field, such as SiNLP: The Simple Natural Language Processing Tool (S. A. Crossley, Allen, Kyle, & McNamara, 2014), TAACO: Tool for the Automatic Analysis of Cohesion (S. A. Crossley, Kyle, & McNamara, 2016), SEANCE: SEntiment Analysis and Cognition Engine (S. A. Crossley, Kyle, & McNamara, 2017), Coh-Metrix (McNamara et al., 2014) and LIWC: Linguistic Inquiry and Word Count[10]. However, this research study applied the LIWC tool for analysing the discourse due to its ability to capture insights from learners' language and its position as a widely-used tool in much language research studies. Although Coh-Metrix has been applied in several studies, this thesis did not utilise the tool due to its inability to perform batch processing through the freely available version.

2.4.1 Linguistic Inquiry and Word Count Tool

The Linguistic Inquiry and Word Count (LIWC) tool, designed by Pennebaker et al. (2001), has been used in thousands of studies to analyse natural language. It gives a quantitative measure to represent user-generated text from several media such as emails, verbal speech, newspaper articles, journaling, poetry, conversations in online chat rooms, and learner-generated text in the education contexts (Stirman & Pennebaker, 2001; L. D. Stone & Pennebaker, 2002; J. W. Pennebaker & Graybeal, 2001; Beevers & Scott, 2001; Mehl & Pennebaker, 2003; Gortner & Pennebaker, 2003; M. L. Newman, Pennebaker, Berry, & Richards, 2003; Rude, Gortner, & Pennebaker, 2004; Alvero et al., 2021). The tool computes the level of usage of different word categories, such as personal pronouns, articles, and affect features. The LIWC tool uses several language categories to capture the social and psychological aspects of the words being used (Tausczik & Pennebaker, 2010). A detailed summary of features included in the LIWC tool is presented in the following section.

2.4.1.1 Summary Dimension

The 'Summary dimension' in LIWC reflects the summary of the overall text through different linguistic measures such as word count, words per sentence, analytical thinking,

[9] https://www.linguisticanalysistools.org/
[10] http://liwc.wpengine.com/

clout, words that are greater than six letters, and dictionary words. Several studies have utilised these summary language cues for analysing a text component (Hartley, Pennebaker, & Fox, 2003). A simple word count and words per sentence of a user-generated text shows how well a user can convert their thoughts and opinions into textual data, representing their verbal fluency. Additionally, the cognitive complexity of a user can also be represented through 'Words per Sentence' (Tausczik & Pennebaker, 2010).

Several research studies (Beaudreau, Storandt, & Strube, 2005; Arguello et al., 2006) have utilised word count and words per sentence in their analysis to convey their importance. In particular, the linguistic complexity of a text can be computed through the average words per sentence and amount of words that are greater than six letters (Arguello et al., 2006). Moreover, dictionary words can measure the non-technical language used by an individual during the discourse, as it measures the percentage of words represented in the dictionary. Therefore, it indirectly computes users' technical language in the text (Tausczik & Pennebaker, 2010).

'Analytical Thinking', as defined by LIWC measures the thinking style of an individual (J. Pennebaker, Booth, Boyd, & Francis, 2015). With usage of words, low analytical thinking suggests informal, personal, and narrative thinking, while higher analytical thinking demonstrates more formal, logical, and hierarchical thinking. The underlying algorithm used to calculate analytical thinking was developed according to the results obtained through the study by Pennebaker et al. (2014). The study (J. W. Pennebaker et al., 2014) uses function words such as articles (a, an, the), prepositions (to, above), personal pronouns (I, her, they), impersonal pronouns (it, thing), common adverbs (so, really, very), auxiliary verbs (is, have), conjunctions (and, but), and negations (no, never) to compute different analytical thinking styles.

The analytical thinking style of discourse has been studied in the prevailing literature. Investigating correlations between language and forecasting skills (Zong, Ritter, & Hovy, 2020), identifying significant differences in employees' language use for brand engagement between high and low ratings on social media (Duncan, Chohan, & Ferreira, 2019), and analysing linguistic styles in different types of celebrities' tweets (Tuan, Aleti, Pallant, & van Laer, 2019) are a few example studies that utilise this language cue. It is important to note, obtaining low scores on analytical thinking does not mean that there is

an intellectual difference between two different individuals who obtained different analytical scores: instead it demonstrates whether or not an individual uses more personal experiences or time-based stories (J. W. Pennebaker et al., 2014).

The linguistic measure 'clout' reflects the level of confidence, certainty, and expertise expressed through discourse. The algorithm behind computing clout from textual data was developed from study results that investigated pronoun usage (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2014). According to the study, people with higher status in a social hierarchy use fewer first-person singular ('I'), and more first-person plural ('we'), and second-person singular pronouns ('you'). Moreover, this status is associated with attentional biases during group interactions where an individual who possesses a higher rank focuses on others, while a lower rank individual focuses on themselves (Kacewicz et al., 2014).

With clout scores, it is possible to understand how well an individual can express their level of confidence and certainty. In other words, they indicate whether individuals are highly confident or uncertain (K. N. Jordan, Sterling, Pennebaker, & Boyd, 2019). According to Pennebaker et al. (2015), a high score suggests high expertise and confidence, whereas a low score reflects a more tentative level. Nevertheless, a low value score in clout does not convey that the information of the given discourse is less accurate. The clout score only indicates the level of confidence of the individual on the given subject (Moore, Yen, & Powers, 2021). It is observed that the use of clout language increased at the last week of the course as learners become more confident and comfortable at the end of the course (Zhu, Herring, & Bonk, 2019).

Apart from the aforementioned studies (Tuan et al., 2019; Duncan et al., 2019), several other research studies have utilised clout as a measure in their language analysis. In an education context, students who obtain a higher clout score refer to others' work in their discourse through their language use such as 'we', or 'you'. Expertise on a topic can also be measured using clout. A recent study (Adaji & Olakanmi, 2019) conducted on Stack Overflow data indicates with time the usage of words such as 'we' and 'you' increases and 'I' decreases in expert groups (i.e., those with a high clout score) while the clout score decreases in non-expert groups, demonstrating the confidence and leadership of the discourse. Similarly, higher clout also indicates students refer to others' work or express how they have related to others' work as they express more 'we' and 'you'

pronouns in their discourse. On the other hand, students who express more 'I' (first person singular) pronouns indicate their work is not social (Oliver, Houchins, Moore, & Wang, 2021). The existing literature shows the linguistic measure clout has been explored in many contexts including mental health (O'dea, Larsen, Batterham, Calear, & Christensen, 2017), social hierarchies (Kacewicz et al., 2014), and education (Smith-Keiling, Swanson, & Dehnbostel, 2018; Smith-Keiling & Hyun, 2019; Oliver et al., 2021; Moore et al., 2021).

2.4.1.2 Function Words

Pronouns, articles, auxiliary verbs, and conjunctions are a few of the categories that are presented under 'Function words' in the LIWC tool. The use of pronouns has been studied frequently to demonstrate the characteristics of a person. The use of first-person singular pronouns is an indicator of self-focus (Pasupathi, 2007). Work by Rude et al. (2004) reveals individuals who are depressed use more first person pronouns (e.g., I, my, me) while expressing their thoughts. Moreover, when expressing their personal experiences, the first person singular pronouns can be observed often (Alexander-Emery, Cohen, & Prensky, 2005). Many research studies have used pronouns to understand language in a successful interaction (Arguello et al., 2006); to understand personality in telling bereavement narratives through language cues (Baddeley & Singer, 2008), word use in emotional narratives (Boals & Klein, 2005; Löckenhoff, Costa Jr, & Lane, 2008), and investigating students' daily social environments and assessing their everyday language (Mehl & Pennebaker, 2003).

Articles, or determiners (a, an, and the) are used to represent the 'givenness' in the discourse (S. A. Crossley et al., 2014). More givenness in a text, indirectly a greater use of determiners, indicates greater text cohesion (Gundel, Hedberg, & Zacharski, 1993).

2.4.1.3 Affect Features

The 'Affective process' in the LIWC tool reflects the emotionality of an individual. It captures the emotional state through words that represent different kinds of emotions. Words such as 'love', 'nice', or 'sweet' capture positive emotions, while 'hurt', 'ugly', or 'nasty' express negative emotions. These measures demonstrate whether a person shows

a positive or negative emotion in their discourse. The role of affect features that capture negative and positive emotions is studied in a variety of contexts, such as analysing the narrative derived from health sectors (Alvarez-Conrad, Zoellner, & Foa, 2001), understanding language in a successful interactions in group communication (Arguello et al., 2006), understanding a personality in telling bereavement narratives through language cues (Baddeley & Singer, 2008), and understanding the language markers in emotional narratives (Gill, French, Gergle, & Oberlander, 2008; Bantum & Owen, 2009). In an educational context, emotions are studied to understand the relationship with subsequent iterations of the course (J. Hu, Dowell, Brooks, & Yan, 2018), to detect verbally abusive behaviours in discussion forums (Joksimovic et al., 2019), or to detect confusion (Atapattu, Falkner, Thilakaratne, Sivaneasharajah, & Jayashanka, 2020) in MOOC posts.

2.4.1.4 Cognitive Process

The 'Cognitive process' in the LIWC tool measures various perspectives, such as insight, causation, certainty, differentiation, discrepancy, and tentativeness (J. W. Pennebaker, Boyd, Jordan, & Blackburn, 2015). Causation words (e.g., 'because', 'effect' and 'hence') and insight words (e.g.,'think', 'know' and 'consider') have been associated with active processing of events and are also expressed at higher levels of emotional and traumatic writing as the individual tries to make explanation statements to explain the difficulty (Tausczik & Pennebaker, 2010). Additionally, tentative words (e.g., 'maybe', 'perhaps', or 'guess') and filler words (e.g., 'blah', 'I mean', or 'you know') are used when an individual is uncertain about a topic (Tausczik & Pennebaker, 2010).

In prior literature, cognitive process has been studied in many discourses from various contexts such as education (Moore et al., 2021), mental health (Friedman et al., 2003) and in emotional narratives (Guastella & Dadds, 2006). In education, the relationship between clout and cognitive processes were explored in MOOC discussion forums. It has been found that discrepancy, certainty and differentiation were significantly negatively associated with clout scores (Moore et al., 2021). Similarly, in mental health (Friedman et al., 2003), cognitive process was studied to understand, how cognitive words in trauma narratives are associated with post-treatment psychopathology (Alvarez-Conrad et al., 2001).

2.4.1.5 Punctuation Marks

Though punctuation marks do not directly represent any direct psychological meaning, they convey informative information about a person's discourse. For example, computing the question marks, and quotations reflects the information-seeking behaviour of a user and the external references in their discourse respectively. The role of punctuation in discourse has been studied previously and it is reported that punctuation can provide informational cues that are vital in understanding the structure of a discourse (Dale, 1991). The broadly studied punctuation in the literature is the 'question mark', used to identify confusion in a discourse and annotating a post as a question (Yang, Wen, Howley, Kraut, & Rose, 2015; Agrawal, Venkatraman, Leonard, & Paepcke, 2015; Z. Zeng, Chaturvedi, & Bhat, 2017; Atapattu, Falkner, Thilakaratne, Sivaneasharajah, & Jayashanka, 2019). The study by Agrawal et al. (2015) depicts learner confusion are strongly correlated with urgency and question variables. Similarly, many research studies have utilised question marks in identifying confusion in MOOC discourse and denoted question marks as one of the top-ranked features for confusion detection (Yang et al., 2015).

2.4.1.6 Time Orientation

Three different time orientations fall under the 'Time Orientation' category of the LIWC tool: past, present, and future orientation. Each orientation is captured through different language cues. In other words, past orientation is captured through past tense verbs and language cues that represent past events. Likewise, present and future orientation are captured through present and future tense verbs respectively. Moreover, these orientations also uses language cues that capture the relevant time orientation. For example words that represent future events, such as 'ahead', and 'hope', exemplify language cues that represent future events. In regards to psychological perspective, people who focus on the past use past orientation language cues in their language. Similarly, people who focus on the future, and who are goal-oriented, use future orientation language cues during their discourse. A recent study by Phillips (2018) investigates the relationship between time orientation words with self-compassion. The study results showed low trait self-compassion individuals used more future oriented words in their discourse. Another study (Kovanović et al., 2018) conducted to investigate students' self-reflections reveals, learners who reflected about previous events highly used past-oriented words in

their discourse. Study also demonstrates past-oriented words are one of the important classification feature in predicting self-reflection.

2.4.1.7 Informal Language

Informal markers in a text are categorised under 'Informal language' in the LIWC tool. It captures language cues such as netspeak language, nonfluencies, filler words, and assent. Netspeak language in the text refers to words that are used in social media and text messages (e.g., emojis, lol). Similarly, filler words and nonfluencies are captured through 'ah', 'mm', and 'you know'. On the other hand, assent language represents agreement through words like 'agree', 'ok', 'yes', and 'okay'. These informal language cues are studied to understand whether there is a significant difference in using this language cue in different group settings (Beaudreau et al., 2005).

2.4.2 Psychological aspects of LIWC Tool

As aforementioned, the design and development of the LIWC tool and its categories is closely associated with psychological aspects. The major intention when creating such a tool is to locate words that are connected with psychology-relevant categories (Tausczik & Pennebaker, 2010). Therefore, it is important to understand the psychological perspective of the LIWC tool. A brief summary of psychological aspects of the LIWC tool, related to this thesis are presented in the section below.

Attentional Focus

A person's attention can disclose much information, such as their priorities, opinions, and how they process. Though content-related keywords explicitly give information about individual attention, it is stated by Tausczik and Pennebaker (2010) that analysing function words (e.g., personal pronouns) can also indicate such attentional focus, whilst its temporal focus can be captured by the use of verb tenses. These function words are used in many research studies to understand cognitive processes and perspectives through language use (Kowalski, 2000; Rude et al., 2004). High usage of first person singular pronouns can be seen in people with depression and emotional pain (Rude et al., 2004). Kowalski (2000) studies pronoun usage while writing about an experience of teasing. It reveals that, depending on the role of the individual (i.e., victim/ perpetrator),

the amount of first person singular and third person pronouns can differ. These examples show that pronouns are a good linguistic cue that indicates an individual's priorities or focus.

Emotionality

It is human nature to respond in extremely different ways to two distinguishable events. Analysing people's emotional response is a way to understand how they are feeling about a given event. Alpers et al. (2005) claim there is concurrent validity in the LIWC tool by examining the correlation between the computer scores of selected categories of LIWC with human ratings. One of their prominent results shows that emotion in a language can be identified by using the LIWC's positive and negative emotion words.

Social Relationships

According to Tausczik and Pennebaker (2010), social processes such as a person's status, cooperation among team members, and the quality of a relationship can be measured by investigating the words being used during a communication process. For instance, a simple word count can identify how well a person is controlling a conversation and also reflects their level of engagement. Assents and positive emotion can shows the degree of agreement.

Thinking Styles and Cognitive Mechanisms

Thinking is an exhaustive and complex process. Thinking styles can be understood through the words people use in their communications. On the other hand, cognitive complexity can be understood by analysing the reasoning: the ability to differentiate among several rival solutions, and the ability to incorporate several solutions (Tetlock, 1981). These reasoning methods are measured by LIWC categories, namely: 'exclusion' and 'conjunctions' respectively. Words like 'but', and 'exclude', which are captured under exclusive words, are able to capture this distinction; whereas words like 'and', and 'also' reflect the ability to incorporate several ideas and also reflect coherent narratives (Graesser, McNamara, Louwerse, & Cai, 2004).

The complexity of a language can be measured by the LIWC tool using measures like cognitive process, prepositions, and words greater than six letters. A topic's complexity and concreteness is reflected through prepositions such as 'with', and 'above' (Tausczik & Pennebaker, 2010). The study by Hartley et al. (2003) shows that, an author uses

more prepositions in the discussion than the introduction or abstract in a journal article since discussions are considered to be the complex part as they incorporate the results and compare them with previous outcomes.

Moreover, causal words such as 'because', and 'hence' and insight words like 'think', and 'consider' can also reflect the cognitive process. Conversely, uncertainty of a topic can be reflected through language cues such as tentative words (e.g. perhaps, or guess) and filler words (e.g. I mean, or blah). The more we incorporate tentative and filler words, the more we convey that information has yet to be processed (Tausczik & Pennebaker, 2010).

2.5 Summary

The prior literature demonstrates the importance of investigating learner discourse in MOOCs. From retaining learners to helping them to navigate through the course with ease, understanding the learner discourse is essential. Discussion forum posts are among the most important visible learner discourse in MOOCs. They have the ability to express the knowledge and thoughts of the learner in an explicit way. This shows the benefits of analysing the discussion forums; hence, this thesis focuses discussion forum posts as its key element of investigation.

To investigate student learning, MOOCs' discussion forums can be approached in multiple ways. However, this thesis selected two different cues: 'learner role' and 'linguistic expressions' for investigations. Therefore, this chapter presents comprehensive related work in two major areas, namely, 'Role Modelling' and 'Linguistic Analysis'.

Several roles have been studied in prior literature, however, studies have emphasised the importance of locating sufficient interactions and collaborations amongst learners and predominantly selected network measures to identify roles in forum discussions. It is also important to note that role categorisation highly depends on the goal of the research and other contextual elements that can influence the collaboration amongst learners. If a learner wishes to discontinue a collaboration or likes to give opinions directly in forums without interactions, applying inappropriate role categories and network measures can mislead. To address the issues found in less structured data, where insufficient post-replies are seen, this thesis identified three different learner roles (information-giver,

information-seeker, and other) using the grounded theory approach to best represent the study data.

Several research studies have used linguistic analysis on MOOC data with various intentions. Therefore, this thesis summarises the related work on the potential of linguistics that can be used to investigate students' learning, along with studies that focus on grade predictions in MOOCs.

Furthermore, this literature review has also analysed the primary language tool, LIWC, used in this thesis to explore the various linguistic markers from learner discourse. It is believed the words we use in our day-to-day life can convey several psychological meanings such as analytical thinking, emotional expressions, and social relationships. Therefore, the study presented the psychological perspective of the LIWC category, which can be later used to infer the linguistic expressions used by different learner groups.

The prior literature demonstrates that roles and linguistic expressions are traditionally studied individually, with different goals. However, this thesis suggests these two features can contribute to investigating student learning. This thesis couples the learner role with linguistic expressions to investigate different aspects of learning, such as through linguistic behaviours and topic contributions of two different learner groups. Chapter 3 presents the research methodology of this thesis and explores different research questions surrounding roles and linguistic expressions that are built upon the findings from related work presented in this chapter.

Chapter 3

Research Design

3.1 Introduction

Understanding students' learning is a vast research area that can be explored according to many different attributes. This thesis attempts to investigate students' learning in Massive Open Online Courses (MOOCs) through learner roles and linguistic expressions. It starts by exploring the learner roles in discussion forums and, subsequently, the investigation extends to understand how the contributions in discussion forums differ across different learner groups, (e.g., students who obtain different course grades) and finally analyses whether students' knowledge and opinions expressed during the learning process have an impact on their final course grade.

Though students leave several traces of their behaviour, this thesis uses the discussion forum data alone to investigate students' learning. To this end, the research focuses on discourse analysis of the opinions, thoughts and knowledge expressed in the discussion forums.

This chapter presents the research design and a high level explanation of the studies conducted in this thesis to discover the relationships between learner roles and linguistic expressions with student learning.

This chapter is organized as follows: Section 3.2 describes the overall research design, Section 3.3 presents the research context and the course environment, Section 3.4 introduces the studies conducted in this thesis, Section 3.5 describes the Machine Learning setup used for the research studies and Section 3.6 provides a summary of this chapter.

3.2 Overall Research Design

The overarching goal of this thesis is to investigate students' learning in MOOCs by extracting features from their language used in discussion forums (i.e., learner roles and linguistic expression). To achieve the aforementioned goal, the design phase of this research study can be divided into four main objectives:

1. Objective 1 - Develop a predictive model to identify learner roles in discussion forum posts.
2. Objective 2 - Explore the relationships between lecture topics and different learner clusters (i.e., High Distinction information-givers, Fail-grade information-givers)
3. Objective 3 - Develop a predictive model to identify linguistic features that contribute to classify learner grades.
4. Objective 4 - Investigate the relationships between linguistic expressions and different learner clusters.
5. Objective 5 - Develop a Linguistic Profile (i.e. linguistic behaviours/characteristic of learners) based on significant linguistic features.

An overview of the research design is presented in Figure 3.1.

This thesis uses quantitative analysis to analyse the data and conclude the findings based on them. Though the obtained data is in text format (i.e., discussion forum posts), the study uses the Linguistic Inquiry Word Count (LIWC) (J. W. Pennebaker, Chung, et al., 2007) tool, which converts these textual data into numeric numbers. Machine learning models were built using these quantitative measures and outcomes were interpreted in accordance with the research goal.

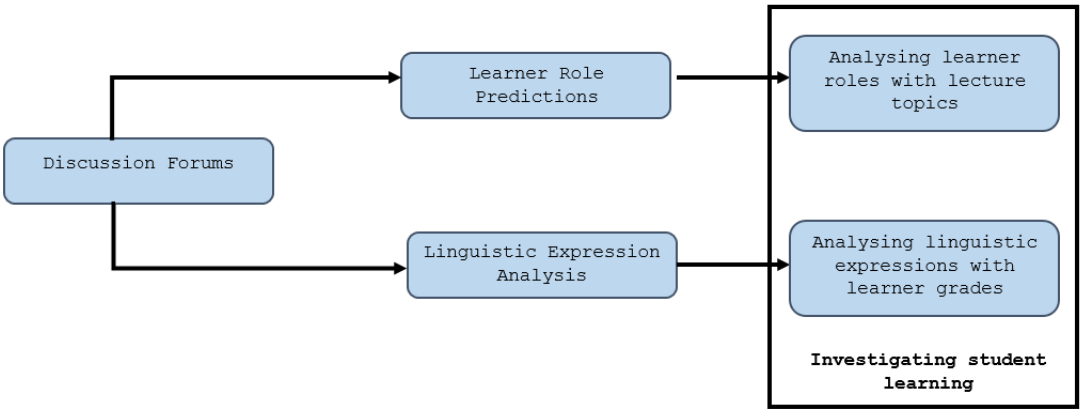


FIGURE 3.1: Overview of the research design

3.3 Research Context

The intention of this section is to describe the research context used in this thesis.

3.3.1 Course environment

The research study was conducted on the "Introduction to Project Management" course offered by AdelaideX on the edX platform in 2016, which is freely available to anyone anywhere around the world. The aim of this course is to explore and understand goals from the beginning of a project, and to consider all the factors that may affect project execution. The course delivers step by step guidelines on how to plan, scope, schedule, cost and manage a project from beginning to end. Since every project relies on the people who are delivering it, the course also enables students to explore how to communicate effectively, manage people and employ leadership skills to deliver a project successfully.

This is a six week course where every Wednesday a new topic is discussed. This course discusses several subtopics in a week using lecture videos. The course consists of lecture materials, knowledge checks and discussion forums throughout the learning process.

The structure of this course is presented below:

- WEEK 1: What is Project Management?
- WEEK 2: The Initiation Phase
- WEEK 3: Understanding the Planning Phase

- WEEK 4: Determining Project Risk
- WEEK 5: Project Teams and Communication
- WEEK 6: Bringing it all together

Each week, an instructor posts a question in a discussion forum where learners can post their responses; however, learner participation is optional. The discussion forum is guided by the following questions.

Week 1: Are you involved in a project?

In this exercise, we would like you to provide an example of a memorable project that you were involved in:

- *Describe what the project was;*
- *Describe the challenges you faced, in particular related to the (4) key elements of all projects.*

Week 2: Applying the process

Now that you have seen the Canvas framework applied to Janet's DIY project, how do you think this could be applied to Peter or Anne's projects? In this exercise we want you to think through, then discuss the initiation process for either Peter or Anne. You may need to review these scenarios again from week 1, then consider the key considerations for the one you choose.

Week 3: Project Planning

As we have seen from the three scenarios presented in week 1, each has its own set of project deliverables. In this exercise we want you to think through, then discuss what you feel are the most important aspects of project planning for each scenario. You may need to review these scenarios again from week 1, then describe the key points for each.

Week 4: Significant Risks

Now that you are familiar with risk management and the PESTLE framework, we want you to identify significant risks. Choose one or more of the course scenarios and identify the most significant areas of risk in the project. For each significant risk identified, suggest

a way it could be controlled or eliminated. You may need to review these scenarios again from week 1.

Week 5: Communication tools

As we have seen, effective project communication is very important to the success of any project. In this exercise we want you to share your favourite communication tools and describe how you use them effectively. If you aren't familiar with this type of technology, then start with those mentioned in the tools and strategies video and investigate further from there.

3.4 Experimental Framework

This section provides an overview of the studies involved in this thesis. Three different studies are performed to investigate the students' learning in MOOCs through discourse analysis. These three studies are presented as separate chapters to address the different objectives of this thesis.

The first study of this thesis focuses on identifying learner roles exhibited by a learner in a given forum post. The next study identifies topics that are discussed by learners in forum posts and is followed by a comprehensive analysis of the relationships between these topics and learner clusters to unveil the correlations among and across them. The final study examines the linguistic expressions in learner posts and presents a set of rules to understand the relationships between the linguistic expressions and learner grades.

3.4.1 Study 1 – Role Modelling

Chapter 5 describes the study performed to identify a learner role in a discussion forum post. The study also involves feature engineering and hyper parameter tuning to identify a machine learning model that outperforms existing models in role identification. This predictive model is implemented, based on the features obtained from feature selection techniques. Subsequently, classification techniques were used in prediction. The study used linguistic-only features to identify the learner roles in discussion forums, disregarding any structural features (e.g., position in the thread, or number of votes). As a result, real-time role identification is possible in a less structured discussion forum. To

ensure the validity of the proposed machine learning model, a cross-course evaluation was performed on the "Risk Management for Projects" course delivered by AdelaideX as it provided rich discussion forums compared with the given range of courses.

This research work is guided by the following research questions (RQ):

- RQ1: To what degree of granularity can a machine learning model predict learner roles in discussion forum posts using linguistic features alone?
- RQ2: What are the linguistic features that contribute significantly towards identifying a learner role that is demonstrated in a forum post?
- RQ3: To what extent can machine learning models that rely on linguistic features be used across courses from similar domains?

3.4.2 Study 2 – Topic Modelling

Chapter 6 presents a study on topic modelling in learner posts to understand the topic's contribution. The primary goal of this study is to investigate in what ways different learner roles (i.e., information givers, or information seekers) and learner clusters (i.e., Pass-grade information givers, or Fail-grade information givers) contribute towards lecture topics. The study analyses the lecture topics discussed by students taking different learner roles. Furthermore, it also investigates how learner roles from different learner grades (i.e., high distinction, distinction, credit, pass and fail) will contribute to lecture topics. For instance, the study seeks to identify different topics discussed between a *'High Distinction information-giver'* versus a *'Fail-Grade information-giver'*. An extensive analysis was performed to identify relationships between the lecture topics and different learner clusters to understand the students' learning patterns. This study contributes to answering the following research questions:

- RQ1: What are the main discussion topics discussed in learner posts?
- RQ2: What are the main discussion topics discussed in different learner clusters?

The results from this study provide insights to support the development of learning strategies and course designs to decrease dropout rates and also help with understanding learners' knowledge achievements.

3.4.3 Study 3 – Linguistic Analysis

Chapter 7 describes a study that focuses on different linguistic features extracted from discussion forum posts. The aim of this study is to extract a set of rules to explain the linguistic behaviour of learners who obtain different grades. This will eventually help to understand the relationship between linguistic expressions that are expressed in learner posts and in different learner clusters. To achieve this aim, the study addresses the following research questions:

- RQ1: How do linguistic features extracted from students' discussion forum posts contribute to learner grade predictions?
- RQ2: What are the significant rules that can be developed using the linguistic features extracted from discussion forums to identify the likelihood of different learner grades?
- RQ3: What are the significant linguistic features that can contribute to developing linguistic profiles of learners?

Outcomes of this study will act as motivations for better course design and easy identification of instructor interventions during the course. Moreover, extracting the significant rules from decision trees will help to provide human-understandable rules that can easily be implemented in a MOOC environment to identify learners who are more likely to obtain a Pass-grade and Fail-grade towards the end of the course.

3.5 Machine Learning Framework

Machine learning is a branch of computational algorithms that are designed to imitate human intelligence by learning from existing data. Fuelled by advancements in algorithms and computer power, machine learning techniques have become powerful tools to discover patterns in data (Biamonte et al., 2017). Machine Learning enables computer systems to make predictions by learning the relationships from previous data.

This section describes the Machine Learning Framework used in this thesis namely: the classification algorithms used to build predictive models (Role and Grade prediction), cost-sensitive learning, the process of stratified cross-validation and evaluation metrics.

3.5.1 Classification Models

Several machine learning classification models were implemented to predict learner roles and final course grades. The classification models used for the predictions are Decision Trees, Logistic Regression, K-Nearest Neighbors, and Ensemble models (XGBoost, AdaBoost and Random Forest). These models are considered to be prominent classification models and have been used to address several research problems in education research.

This thesis used 'Entropy' as the tree node splitting criterion for the tree-based models. Ensemble learning models are constructed by combining multiple predictive models to improve accuracy. Bagging and Boosting are well known ensemble techniques used in predictive models.

3.5.2 Cost-Sensitive Learning

In real-world data, expecting a balance distribution among different classes is impossible. Such imbalances between different classes need to be treated in the training data to avoid biased and inaccurate classification models. This thesis uses cost-sensitive learning by mapping class weights inversely proportional to class frequencies.

The weights were calculated using the scikit-learn [1] class weight method as defined below:

$$\frac{n_{\text{samples}}}{(n_{\text{classes}} * [n_1, n_2, ...])} \tag{3.1}$$

In the above equation $n_{samples}$ is the number of instances in the training sample, $n_{classes}$ is the number of classes and n_1,n_2,... are the number of instances in each class (Elkins, Freitas, & Sanz, 2019).

3.5.3 Stratified Cross-Validation

Cross-validation is a technique used to evaluate predictive models by partitioning the original data into k equal size disjoint subsets (folds). During model training, a classifier is constructed with k-1 subsets (i.e., the training data). The remaining subset (i.e., the test data) is used for testing the model. This cross-validation process is repeated until

[1] https://scikit-learn.org/stable/

all the k subsets have been used as test data (X. Zeng & Martinez, 2000). This ensures every instance in the data set becomes a testing instance in one of the iterations and each instance is tested only once. Figure 3.2 illustrates the process of cross-validation with 10-fold cross validation (i.e., k=10).

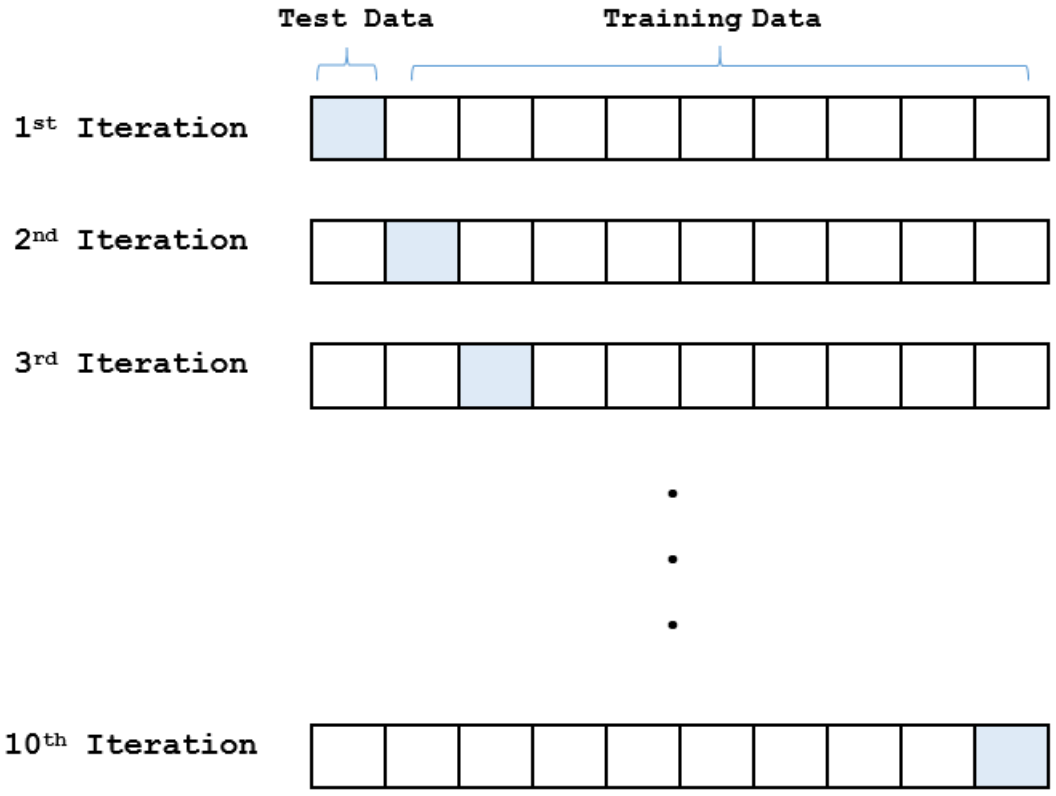


FIGURE 3.2: Cross Validation (k=10 fold)

In k-fold Stratified Cross-Validation, a data set D is partitioned into k folds in such a way that each class is uniformly distributed among the k folds. This ensures the class distribution in each fold reflects the original data. In regular cross-validation, a specific class could be unevenly distributed, where some folds contain more instances of a particular class than another. Stratified cross-validation eliminates such distortion in class distribution, which is usually caused by regular cross-validation.

3.5.4 Evaluation Methods

Assessing the quality of predictive models is an essential task during model development. Several evaluation metrics exist to evaluate the performance of different machine learning

algorithms. These metrics help to examine the effectiveness of predictive models and their predicting capabilities.

The performance metrics used to evaluate the performance of machine learning models are based on a confusion matrix. A confusion matrix for binary classifiers is presented in Figure 3.3.

		Actual Class	
		Positive	Negative
Predicted Class	Positive	**TP**	**FP**
	Negative	**FN**	**TN**

FIGURE 3.3: Confusion Matrix

In a confusion matrix, columns represent actual classes; whereas rows represent the predicted classes. The symbols that are denoted in Figure 3.3 are described below:

- True positives (TP) - Number of instances that are correctly predicted as positives.
- False Negatives (FN) - Number of positive instances that are predicted as negative.
- True Negative (TN)- Number of instances that are correctly predicted as negatives.
- False Positive (FP) - Number of negative instances that are predicted as positives.

This thesis used the following measures to evaluate the performance of classification models:

Precision - It measure the proportion of predicted positive cases that are correctly real positives (Powers, 2020).

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{3.2}$$

Recall/Sensitivity - It measures the proportion of real positive cases that are correctly predicted positive (Powers, 2020).

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3.3}$$

F1 score - It measures the harmonic mean between precision and recall (Powers, 2020).

$$F1score = 2 * \frac{precision * recall}{precision + recall} \tag{3.4}$$

Area Under the ROC Curve (AUC)

The ROC (Receiver Operating Characteristic) curve plots the true positive rate (sensitivity) against the false positive rate (1-Specificity) at numerous classification threshold points (Powers, 2020). A sensitivity/specificity pair corresponding to a particular decision threshold is represented by each point on the ROC curve. The AUC computes the area under the ROC curve, in particular, it measures the entire two-dimensional area under the ROC curve. AUC provides an aggregate measure of performance across all possible classification thresholds.

3.6 Summary

This chapter outlines a high-level view of the research methodology and the studies involved in this thesis. The research methodology outlines how learner roles and linguistic expressions are used to investigate the student learning in MOOCs. Subsequently, the research context of this thesis is comprehensively described to understand the learning environment of the research data. The study uses discussion forum data obtained from the 'Introduction to Project Management Course' provided by the AdelaideX platform. The course structure and the format of the discussion forum have been discussed to describe and analyse the learning context.

Additionally, this chapter described the three studies involved in this thesis. Each study presented a set of research questions to achieve the aim of the study. Each of these studies is explained separately in Chapters 5, 6 and 7, respectively. These chapters separately discuss the study methodology, results and present a discussion to address the research

questions. Finally, this chapter explained the machine learning set up used in this study, along with the evaluation metrics. The next chapter presents an overview of the dataset and a comprehensive analysis of the research data used in this thesis.

Chapter 4

Research Data

4.1 Introduction

This chapter presents the overview of the research data used in this thesis and provides comprehensive details on data collection and data annotation. This dataset has been used as the main data source in each research study identified in Chapter 3.

Chapter 2 reviewed the existing learner roles identified in the literature. A role in forum data purely depends on the personal (e.g., social personality) and contextual elements (e.g., optional participation, or frequent instructor intervention). For example, a student who is an extrovert will likely create a network among peers, whereas another student might only post responses to the given question in a forum thread. These elements (i.e., personal and contextual) can highly influence the emergence of roles in forums. As the literature (Strijbos & Weinberger, 2010; Dowell & Poquet, 2021) emphasises, different sets of roles can be seen across different forum data. Furthermore, the researchers' intention or perspective to investigate the discussion forum will be different from one course to another. This shows there are no universal role categories identified in the research literature to suit every discussion forum. Therefore, it is important to analyse the Massive Open Online Course (MOOC) forum data and alter the role categories to suit the given research data and context.

The discussion forum data obtained from AdelaideX courses was meticulously analysed and eventually learner roles were identified using a grounded theory approach (Vollstedt & Rezat, 2019). Discussion forum posts were annotated by two independent annotators,

according to the roles identified during the grounded theory approach. These identified learner roles can be applied in any (structured or less structured) learning environment as these roles are identified at post-level. The intention of this study is to identify the learner role given a forum post. The term learner role is used instead of post role as this study identifies roles exhibited by a learner at a given point of time (i.e., post).

This chapter is organized as follows: Section 4.2 presents the overview of the data, Section 4.3 introduces the grounded theory approach used to identify the learner roles, Section 4.4 presents the information on the AdelaideX dataset, followed by data sampling and data pre-processing in Section 4.5 and Section 4.6 respectively. In accordance with the leaner roles identified in Section 4.3, Section 4.7 explains the data annotation process. This section also discusses the disagreements observed during the data annotation and provides a statistics on learner roles. Finally Section 4.8 provides the summary of this chapter.

4.2 Overview

The goal of this thesis is to investigate students' learning in MOOCs through learner roles and linguistic expressions. In doing so, this thesis focuses only on discussion forum data to identify the roles and linguistic expressions that are expressed via learner posts. Two different courses, namely 'Introduction to Project Management' and 'Risk Management for Projects' were selected from the AdelaideX[1] learning platform as the data source for both training and evaluation. As this research is formulated around discussion forums, these two courses were selected due to the richness of their discussion forums, compared with other courses offered by AdelaideX.

Furthermore, this research study includes analysis of not only learner roles but also learner grades, topic identification and rule extraction, which are beyond identifying roles in forum posts. During model building, for example, experiments like topic modelling required an additional consecutive semester's data to train the topic model. Therefore, the study used the AdelaideX platform from which consecutive semester data can easily be retrieved.

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

According to the literature (Strijbos & Weinberger, 2010; Dowell & Poquet, 2021), different roles have been identified in Computer-supported Collaborative Learning (CSCL) such as 'Pillar', 'Generator', 'Hanger-on', 'Lurker', 'Captain', 'Over-rider', 'Free-rider', 'Ghost', 'Follower', 'Socially Detached', 'Influential Actor' and 'Hyper Posters'. However, it is important to analyse the dataset obtained from AdelaideX before annotating the roles of each forum post to determine which roles emerge in this context.

One of the motivations of this study is to enable real-time role identification without any delays and also to identify a set of roles to overcome the limitations (i.e., observed limited peer networking) that reside in the dataset. The analysis of the study data reveals that among comment-thread initiators (i.e., learners who initiate a thread), 65% of the learners do not contribute in any post-replies (i.e., learners have not posted any reply comments to other threads). In other words, they have only initiated a thread. This shows that formulating a social network analysis might classify them as 'Socially Detached' learners where high participation can be seen in these types of learners with ineffective or surface-level contribution to a group conversation (Dowell & Poquet, 2021). Moreover, their discourse is less likely to respond to others during the learning process. Dowell and Poquet (2021) construct directed and weighted social networks, where weights that characterise the replying tendency of a learner (Chen & Poquet, 2020) were given by calculating the number of replies sent from source to target. Thus, adapting a similar social network in this dataset will lead to inefficiencies as most of the learners were identified with zero post-replies.

This thesis also reviewed the process and datasets that applied Social Network Analysis (SNA) to predicting the roles. According to Marcos-Garcia et al. (2015), the study suggests role identification using an SNA method. The study proposes 'student-animator' as one of the roles. The role is defined as containing active students who are motivating their peers. The study identifies 'student-animator' using the SNA indexes, where their outcentralisation (outdegree, or outcloseness) index will result in high value. The study measures the SNA index of a particular student and compares them with the statistical properties (minimum, maximum and average) of the SNA indexes with the remaining class to determine the role.

The study by Dowell and Poquet (2021) applies SNA to the dataset obtained from 'Instructional Methods in Health Professions Education' course, delivered by Midwest

University on the Coursera platform. Their data description emphasises the importance of peer-interactions. They describe their dataset as follows:

"It is important to note that social learning was among the learning objectives the instructor team set out for this particular MOOC. Towards this goal, the discussion forums were specifically crafted to engage learners in reflection and information sharing in response to the content of the Module/Section at hand. Some are focused on the learner and their perceptions and self-reflections. Others ask the learners to share information for their best-practices and/or experiences and engage in substantive discussion. As such, this MOOC is an appropriately suited context to explore learner profiles and forum discourse dynamics in a scaled learning environment."

In an online education setting, collaborative learning can be seen in both 'closed, credit-based' online learning courses and MOOCs, nevertheless learning with peers will be different in each learning setting. The closed, credit-based online learning course replicates the traditional face-to-face classroom environment. This course environment is instructor-driven, whereby the instructor controls and structures the discussion, and it results in well-structured discussion forums. These courses promote collaborative learning, and group work is expected to contribute to the final grading scheme. Meanwhile, group formation in MOOCs is often not a course requirement; they are impulsive, and not regulated by instructors. Given the massive number of learners, collaborative learning in a MOOC is often student-driven, chaotic and optional.

As a result, it is apparent that applying these existing emergent learner roles of CSCL in a less structured MOOC setting is impossible due to the inability to capture the required level of communication among peers. Furthermore, the characteristics of the existing learner roles cannot be observed in such MOOC settings. For example, roles like Captain and Pillar require a further level of comments to identify them. Identifying learner roles in a less structured MOOC setting is still vital, to diminish the impacts (i.e., high chance of course attrition, confusion, and misconception) caused by a lack of understanding of course content. Due to the aforementioned limitations, and to promote a real-time role identification without delay, this thesis requires a set of roles that suits the given study data. Therefore, this thesis intends to identify learner roles based on a grounded theory approach.

4.3 Grounded theory approach

Grounded theory is a methodology that uses an iterative process for theory development. With grounded theory, theories and concepts emerge that are grounded in the empirical data. The iterative process entails data collection, data analysis and theory development. This iterative process is continued until new data do not contribute to a considerable amount of theory development (Vollstedt & Rezat, 2019). According to the given literature (Vollstedt & Rezat, 2019), grounded theory supports different coding techniques (Glaser, 1978; Mey & Mruck, 2011; Strauss & Corbin, 1990; Teppo, 2015) to code the captured data. Following Strauss and Corbin (Strauss & Corbin, 1990), coding techniques can be categorised as follows: open, axial, and selective coding. Primarily, the study follows an open coding procedure and highlights categories that reside in the data through rigorous data analysis. During the grounded theory approach, open coding is generally the first approach to data analysis. Open coding focuses on categorisation and conceptualisation of data where it categorises the data into discrete parts. This thesis uses open coding as a first step to analyse the data with the intention to develop a coding framework that best describes each sub-group in the data. Initially, pilot study data was analysed and categorised into several categories including: questions, report of an issue, answer, resource recommendations, self-introduction, and thanking posts.

TABLE 4.1: Initial categories with examples

Post types	Example Posts
Questions	How can we implement risk management in the organisation?
Report of an issue	Dear AdelaideX Team Now I'm able to View the Certificate, but unable to Download it. Please provide me the Download privileges.
Answer	Risk identification in my Organisation was being performed with less attention to details except areas relating to finance but with PESTLE we can now work out an orderly and more detailed structure to risk identification in various sectors of the company.

Resource recommendations	You can find more about the risk management in the below article https://www.smallbusiness.wa.gov.au/business-advice/insurance-and-risk-management/risk-management
Self-introduction	Hi, I have been in process and/or project management for 7 years and truly love the field. I have not taken any other courses here so I am happy to see that you found MOOC to be enjoyable. Good luck and I look forward to learning alongside of you.
Thanking posts	Thanks a lot for this response, i got your point. Valid! Thanks !

As the primary objective of this study is to identify learner roles and explore the interconnection between roles and their learning patterns, the study followed 'axial coding' to identify potential roles. Strauss and Corbin (Strauss & Corbin, 1990) define axial coding as investigating the associations between categories that have been developed during the open coding process. Following axial coding, the study examines the connection between the codes identified in the open coding. With the intention of establishing real-time role identification and being able to apply the roles to less structured online courses, the study identifies the roles as 'information-givers' (IG), 'information-seekers' (IS) and 'others' (O), similar to the previous study (Hecking et al., 2017). These categorisations help to understand those learners who are seeking and giving information. Moreover, these identified roles can be applied to both large and small scale group settings. For instance, in a small group collaborative learning environment, a Captain can ask questions or provide answers to keep the group on track. An instructor cannot pay less attention to students who are Captains in the group communication. Attention needs to be given to Captains who seek information as they can also be in a confused mode on a particular lecture topic. This situation needs to be looked with immediate concern as will any other roles who seek information. Therefore, it is important to detect the seekers irrespective of their roles, as aforementioned in the literature (i.e., Captain or Socially Detached).

Examples for each role are presented below.

1. Information-seeker
    - 'How can we implement risk management in the organisation?'
    - 'Can someone from the staff update on our certificates.'
2. Information-giver
    - 'It is important for stakeholders to be involved in the risk management process - from identification through to response as they will have knowledge of risks in their area of responsibility that project managers may not be fully aware of.'
    - 'According to my dashboard, certificate will be available in August'
3. Other
    - 'Hi <participant> from Australia!'
    - 'Thank you for the reply'

The study has created one annotated dataset in particular to facilitate machine learning and language models in order to better understand the learners' behaviours in a less structured learning context. This data is appropriate for researchers and academics (i.e., instructors and lectures), who are interested in learners' behaviours in a less structured learning environment. The dataset described below captures the detailed information of the interactions happening between learners in two MOOCs, along with the grades obtained by learners. The primary use of this dataset is to facilitate the role identification in CSCL in real-time using the given discourse.

4.4 The MOOC Posts Dataset

The thesis uses discussion forum posts from two of the courses provided by the AdelaideX[2] learning platform as the data source. AdelaideX is an online learning program provided by The University of Adelaide with a range of free online courses delivered on the edX platform. The study selected 'Introduction to Project Management' and 'Risk

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

Management for Projects' as they provide rich discussion forums compared with the given range of courses. The context of these courses is described below.

Introduction to Project Management is an introductory course on project management that delivers essential project management knowledge and skills. There is no prerequisite required to enroll for the course. The course is six weeks in length and the weekly workload is 2-3 hours. Course materials include lecture videos and course handouts. Assessment needs to be completed to obtain a certificate. There is a dedicated discussion forum where learners are expected to post at least once per discussion activity and invited to respond to their peers.

Risk Management for Projects is an introductory level course with no prerequisites. The objective of this course is to deliver the fundamentals of risk management and its applicability in real world settings. The course has five weeks' duration and the workload is 2–3 hours per week. The course includes lecture videos and course handouts. To qualify for a certificate, question activities and assignments must be completed. The course provides a discussion forum to interact with peers and instructors. Learners are invited to post at least one post per discussion activity.

The discussion forum data derived from the above courses is directly delivered as JSON (JavaScript Object Notation) format, which is later converted into command-separated values for analysis. The major two data files obtained from the AdelaideX are as follows:

1. AdelaideX_⟨coursename⟩_discussions
2. AdelaideX_⟨coursename⟩_coursegrade

Table 4.2 presents the main attributes in the data file.

Figure 4.1 shows the three levels of communication that can be seen in the discussion forums of edX platforms, which are stored in two types of objects in the discussion forum post data file.

- A Comment Thread (first level of interaction): A post that initiates a new thread.
- A Comment (second and third levels of interaction): A direct reply given to the comment thread is considered as a comment. Any further responses given to a reply are also stored in Comment objects.

TABLE 4.2: List of attributes appears in data file

Attributes	Description
`Author_id`(de-identified)	Unique identification number for learners.
Body	Text in learner post. UTF-8 encoded.
`Comment_thread_id`	Unique identifier of a thread.
`Created_at`	Time comment was posted. Timestamp in UTC.
Title	Title of the thread. UTF-8 string.
`Comment_id`	Unique identifier of a comment

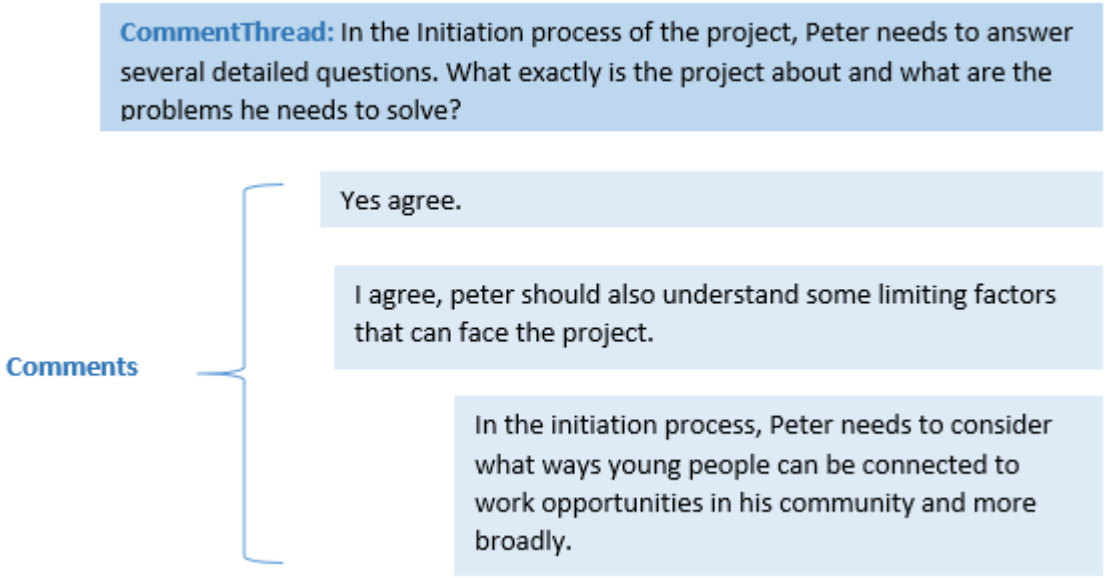


FIGURE 4.1: Discussion forum structure - Comment Thread and Comment Object in the AdelaideX discussion forum

4.5 Data Sampling

The study examined discussion forum posts from the 'Introduction to Project Management' and 'Risk Management for Projects' courses offered in 2016 and 2017 respectively. A total of 9,497 (Project Management - 8,300 user posts, Risk Management - 1,197 user posts) user posts were extracted from 1096 (Project Management - 982 users, Risk Management - 114 users) different users for data annotation. To extract useful and accurate insights from machine learning models, it is important to filter those students who have frequently engaged in the discussion forum. In doing so, learners who have posted more than six posts were selected in 'Introduction to Project Management' as the entire

semester runs for six weeks. Similarly, in 'Risk Management for Projects', the study chose learners with more than five posts due to its five weeks' duration.

The data analysis revealed, an average of 488 user posts were posted each week in the Project Management course with an average of eight discussion posts were posted by each student. Similarly, an average of 149 user posts were posted each week in the Risk Management course. Furthermore, an average of ten user posts were posted by each student.

4.6 Data Pre-processing

The first step is to pre-process the input file before conducting the studies. The input data source is a csv (comma separated values) file, which is extracted from a json object file. These files were preceded by several pre-processing steps as follows:

1. Removal of non-English posts - Few non-English posts were observed while preparing the dataset for the analysis. Therefore, 63 non-English posts were removed from the final data set before proceeding with the experiments.

2. Removal of non-informative words such as stop words (e.g., 'of', or 'is') - Stop words are high frequency words that appear in a text but do not provide much meaning to the given text. Thus, these non-informative words have been removed. However, the stops words included in the LIWC dictionary were not removed from the analysis.

3. Lemmatisation - The study used the lemmatisation technique to derive a common base form for the words used in a given discourse. This helps to minimise the vocabulary space used for the analysis. For instance: 'play', 'playing' and 'plays' were collapsed to the base form 'play'.

   However, studies (Dale, 1991; Jones, 1995) advise punctuation has potential to add meaning in discourse analysis, thus the study decided to retain that punctuation which can contribute in deciding role identification.

4.7 Annotation

After retrieving a set of role categories using the grounded theory approach, the study data was annotated for model training and for further analysis. The purpose of the data annotation is to label the discussion forum posts with meaningful classes. Two independent human evaluators were chosen to annotate the forum posts manually, as follows:

- Information-seeker (IS)
- Information-giver (IG)
- Other (O)

Initially, a pilot study was conducted with 500 posts to familiarise and understand each learner role category. Annotator training was conducted on a sample dataset where 100 posts were given at a time to train them until the annotation process was stable. The annotators were asked to highlight the words that enabled them to assign a learner role for each discussion forum post. Posts that were not certain about the learner role were manually analysed and discussed until agreement was reached.

Afterwards, the original study data set with 9,497 user posts was given to the annotators in four iterations. The annotation differences were discussed and resolved before moving to the next iteration. The first iteration consisted of the 'Risk Management for Projects' discussion forum posts and other three iterations contained 'Introduction to Project Management' discussion forum posts. The number of posts provided for each course with their respective iterations are shown in Table 4.3.

TABLE 4.3: Number of posts in each iteration

Course Name	Iteration	Number of Posts
Risk Management for Projects	Iteration 1	1197
Introduction to Project Management	Iteration 2	3000
	Iteration 3	3000
	Iteration 4	2300

TABLE 4.4: Guidelines for learner role annotation

Learner Role	Description	Example Posts
Information-Seeker - A person who seeks for data/ information	**Question** - Posts that seek information regarding the course.	How can we implement risk management in the organization?
	Clarification requests - A request that is made to clarify.	Can you please explain in brief about all three diagramming techniques? i am confuse in it.
	Logistical - Posts that seek information regarding the availability of course materials, issues with grading.	Please could you provide copies of the presentation videos. Could not download the videos. I would like to know if there other ways of making payment for my Certificate? Hello. Please can anyone tell me how to download the texts? I can download the videos, but not the Texts. Any suggestion? I have upgraded to verified 1 day before the deadline, I get the message "Your certificate was not issued because you do not have a current verified identity with edX. Verify your identity now." However, I have already verified my identity. Can you check that my status is ok and issue the certificate?

	Report of an issue/Technical support -These posts seek technical support about course website issues, or compatibility issues.	GANNT chart doesn't work in Firefox web browser. I accessed the 2nd module (Risk Identification) in my laptop. I wasn't able to finish it. Now, I wanted to continue in my mobile phone but I can't find the 2nd module. If accessed in one platform the module can't be accessed another platform anymore?
Information-Giver - A person who provides data/ information.	**Answers** - These posts give information regarding the course.	Risk identification in my Organisation was being performed with less attention to details except areas relating to finance but with PESTLE we can now work out an orderly and more detailed structure to risk identification in various sectors of the company.
	Resources recommendations - Provide resources outside of the course materials.	You can find more about the risk management in the below article https://www.smallbusiness.wa.gov.au/business-advice/insurance-and-risk-management/risk-management
	Logistical - Posts that give information regarding the availability of course materials, and issues with grading.	I think certificates are only issued after the Edx course completion date. That's what happened on the last course I did.

Other - Posts that do not fall under the above two categories.	**Social/Affective** - Posts that contain social and affective aspects. Social aspects are related to posts that contain self-introduction, or greetings. Posts that reflect emotions (e.g., gratitude) fall under affective.	I am really grateful to the Project101 team. It was a great pleasure for me to receive the lessons from you.I think they was just as useful as they is interesting. Thank you, it is really helpful. Hey everyone, Greetings! I am an Engineer with major in Computer Science and Technology and I am pursuing this course to enhance my knowledge in multiple sectors.

Annotators were given detailed guidelines with examples for each learner role (presented in Table 4.4) that were identified during the grounded theory approach. These guidelines were presented in accordance with the existing work carried out by Hecking et al. (2017) and Wise et al. (2017). Annotators were requested to select a learner role along with the keywords that insisted on choosing a learner role represented in the overall text.

According to Cohen's kappa, a high inter-rater agreement (k= 0.924) between the two annotators ensures the validity of the human annotation.

4.7.1 Resolving Disagreements

During data annotation, major disagreements were observed when a forum post contained more than one learner role. These kinds of post can be categorised into two scenarios as follows:

- Scenario 1: A learner role dominates over another learner role – During these situations, dominant learner roles are assigned for the discussion forum posts. The example below demonstrates two learner roles. The first half falls into the 'other'

category, which then is converted into 'information-seeker'. Even though it contains both learner roles, the information-seeking in the text dominates the other.

"Thanks for the course and all. I learnt alot but if you could just provide a link to download the certificate as a pdf file ,it will be appreciated as participants will agree to that. Thanks edX team.".

- Scenario 2: More than one role with equal significance – These kinds of learner posts are discussed among the annotators and agreed accordingly. The example below exhibits two learner roles: information giver and information seeker respectively.

  "Risk management in my workplace got off in response to a revised legislation to incorporate that practice into every public organization's operations. In its infancy, it takes a lot of effort to check those boxes and making sure documentations are in place. It used to be if any negative risk reared its effect, it's just part of the so-called Murphy's Law and dealt with promptly only to have the same issue - in similar form - popping up the next time. Too idealistic? We seem to be making headway thanks, partly, to the size of the team. What about you out there? Big corporations? Bureaucratic curmudgeon? I'd like to know."

  The above learner post is considered as an information-seeker. Even though it provides information, learner tries to seek information which needs to be addressed first when identifying the learner role.

4.7.2 Annotation Statistics

The overall statistics of the data analysis in both AdelaideX courses shows the majority of the learner posts are from information-givers, at >50%; whereas information-seekers fall between 10% and 13% in the entire dataset. Nearly one third of learners fall under the 'other' category. There is a large disparity in learner role distribution due to our aforementioned data selection method (i.e., data is selected to understand the learning, which results in capturing whole data posts from each student). The study did not capture the entire comments that come under a single comment thread. Therefore, there is a possibility to miss the information-seeking posts that have been made under initialisation threads or information-giving posts. To avoid class-imbalance in the data, class weights were used during the machine learning modelling as presented in Section 3.5.

4.8 Summary

This chapter presents a manually annotated, carefully curated data set for role prediction. This thesis identifies learner roles as one of the attributes to investigate students' learning, demonstrating the importance of identifying learner roles that suit the research data. In doing so, the grounded theory approach is used to identify learner roles to best represent the given dataset.

This chapter also presents a comprehensive annotation process that can be adapted in a similar role identification problem. The next chapter presents the first study of this thesis, 'Role Modelling'. It presents a predictive model that uses the annotated research data described in this chapter to identify learner roles.

Chapter 5

Role Modelling

5.1 Introduction

Chapter 3 and Chapter 4 presented the research methodology and research data respectively. Using the collected student forum data, this chapter presents the very first study in this dissertation to investigate the learner role within a discussion forum post. According to the literature, identifying learner roles in discussion forum posts completely depends on the learning environment. Since the learning contexts differ from one another, a learner role in a well collaborated educational forum cannot be applied to a less structured education forum. This emphasises the importance of identifying learner roles in accordance with the discussion forum context.

While prior literature (Strijbos & De Laat, 2010; Dowell & Poquet, 2021) classifies user roles such as 'Captain', Followers, 'Socially Detached' and 'Influential Actors', this thesis identifies learner roles as 'information-givers', 'information-seekers' and 'others' using the grounded theory approach (Vollstedt & Rezat, 2019). This type of role categorisation can be beneficial not only in the education domain but also in other domains to categorise the user roles as information-givers and information-seekers. For instance, this type of categorisation helps easy detection of the seeking behaviour of users who post in medical forums (e.g., cancer forums) and helps to provide timely support for users who need assistance for their medical conditions (Mayer et al., 2007; Ofran, Paltiel, Pelleg, Rowe, & Yom-Tov, 2012).

The main objective of this chapter is to identify the learner roles in discussion forum posts using only those linguistic features that are extracted from the learners' discourse. Discourse features enable real-time role identification as they do not incorporate any structural elements such as votes, views and thread positions. To address this objective, this study is guided by the following research questions:

- RQ1: To what degree of granularity can a machine learning model predict learner roles in discussion forum posts using linguistic features alone?
- RQ2: What are the linguistic features that contribute significantly towards identifying a learner role that is demonstrated in a forum post?
- RQ3: To what extent can machine learning models that rely on linguistic features be used across courses from similar domains?

This chapter presents a machine learning model that automate the classification of these learner roles in discussion forums posts. The key findings of this study indicate that learner role identification is possible using the discourse features alone. Moreover, cross-course evaluation shows that the machine learning model used in this study can be used to predict the learner roles in discussion forum posts that are extracted from similar courses.

This chapter is organised as follows: Section 5.2 provides the detailed research methodology used in this study, including feature extraction, feature engineering and hyper-parameter tuning. Section 5.3 outlines the setup used in this experiment. The key findings of the study that contains the model building and model evaluation are presented in Section 5.4. Section 5.5 discusses the major findings of the study and its practical implications. Finally a summary of the chapter is presented in Section 5.6.

5.2 Methodology

The aim of this chapter is to identify the learner roles in discussion forum posts using the linguistic features that are extracted from the learner discourse. Discussion forum data obtained from the 'Introduction to Project Management' course and 'Risk Management for Projects' course offered in 2016 by the AdelaideX learning platform were used in

this study. Using the grounded theory approach (Vollstedt & Rezat, 2019), the study identified the learner roles as information-giver, information-seeker and other. These roles are identified at post-level. The terminology 'learner role' is used instead of post role as this study identifies roles exhibited by a learner at a given point of time (i.e., post).

After identifying the learner roles that best represent the study data, data annotation was performed, as mentioned in section 4.7. Subsequently, a predictive model was built to identify the learner roles in forum posts. The following steps were executed prior to building the predictive model:

1. Feature extraction
2. Feature engineering
3. Hyper-parameter tuning

Furthermore, the normality of the data distribution (i.e., linguistic features) was tested to determine the statistical test needed to identify the significant differences between learner roles. It is vital to perform a normality test, as it is one of the underlying assumptions of much statistical analysis. There are several normality tests prevailing in the literature, such as the Shapiro–Wilk and the Kolmogorov–Smirnov tests, to test the normality of the distribution (Öztuna, Elhan, & Tüccar, 2006).

According to Yap et al. (2011), the Shapiro–Wilk test is the most powerful test used for asymmetric distribution. As a result, the normality test for each linguistic feature (dependent variable) was performed using the Shapiro–Wilk test. The test results confirm that the linguistic features are not normally distributed. Therefore, the study selected a non-parametric test to evaluate the significant features.

An overview of the methodology to identify a learner role is presented in Figure 5.1. The details of each step are presented in the following sections.

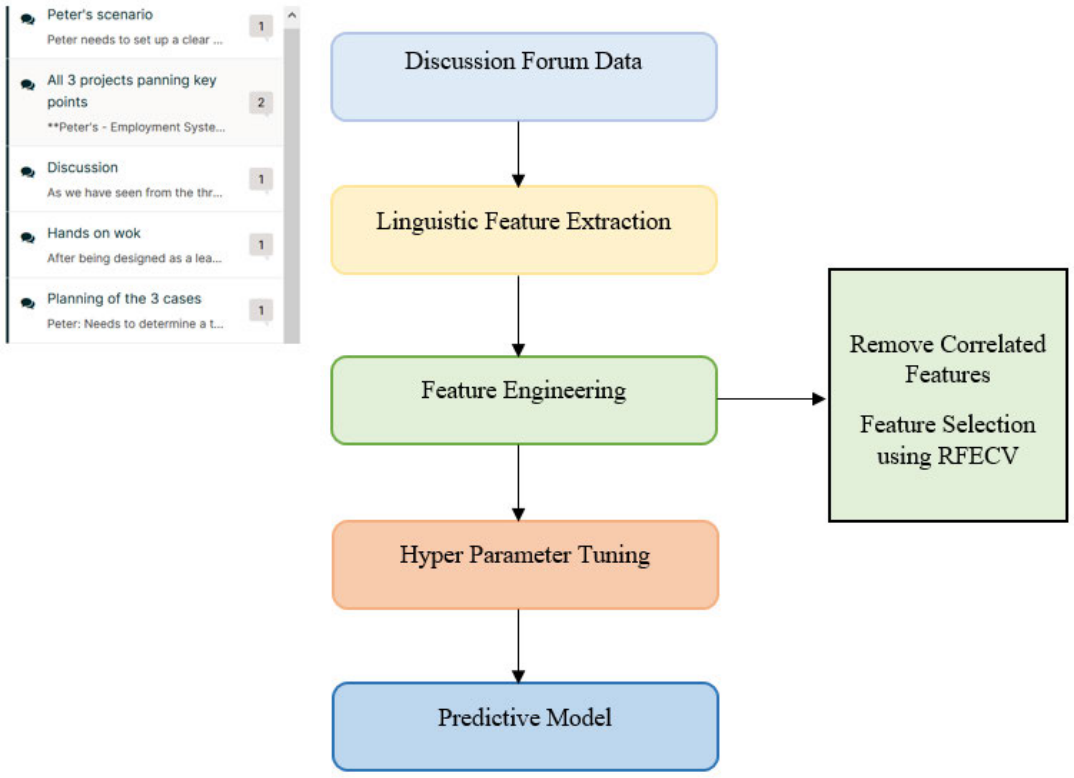


FIGURE 5.1: Overview of the methodology

5.2.1 Feature Extraction

With the intention to study the diverse cognitive, emotional and structural elements present in learner written discussion forum posts, the Linguistic Inquiry and Word Count (LIWC) tool was selected to perform feature extraction. Linguistic features from discussion forum posts were extracted using the Linguistic Inquiry and Word Count (LIWC2015) (J. W. Pennebaker, Boyd, et al., 2015) tool. LIWC is a text analysis tool that has been used widely in several pieces of text analysis research (Arguello & Shaffer, 2015; Wen, Yang, & Rosé, 2014; Litvinova & Litvinova, 2018; Andy, Andy, et al., 2021). LIWC2015 (J. W. Pennebaker, Boyd, et al., 2015) is the latest version, with significant modifications to the existing dictionaries used in the previous versions, LIWC2001 (J. W. Pennebaker et al., 2001) and LIWC2007 (J. W. Pennebaker, Booth, & Francis, 2007).

LIWC has a processing component and dictionaries that take text as an input and process and compare the text with a predefined dictionary to provide numeric values for various linguistic features (Tausczik & Pennebaker, 2010). This predefined dictionary

consists of several categories and each category is a collection of words that describe the category. The LIWC2015 (J. W. Pennebaker, Boyd, et al., 2015) default dictionary consists of approximately 6,400 words, word stems and select emoticons. Due to its ability to analyse numbers, punctuation, and short phrases, "netspeak" language can also be included in text analysis (e.g., b4, :) is coded as a preposition and positive emotion respectively).

The LIWC tool extracts numerous psychological and linguistic features from text, written in natural language. Empirical evidence (Gottschalk & Gleser, 1979; Weintraub, 1989; P. J. Stone, Dunphy, & Smith, 1966) shows that there is an association between language and the state of mind: specifically that psychological aspects are involved in the use of language. LIWC provides an effective and efficient method for studying these psycholinguistic perspectives. A comprehensive explanation of the psychological aspects of the LIWC tool is presented in section 2.4.

An overview of the LIWC feature space that was used to classify information-seekers/ information-givers/ others is described below:

- Linguistic Processes (e.g., summary dimensions, function words)
- Grammar Features (e.g., interrogatives)
- Psychological Processes (e.g., affective processes, cognitive processes and time orientations)
- Punctuation (e.g., question marks)

5.2.2 Feature Engineering

Feature selection is a technique that selects a subset of the most significant and informative features from a given dataset to improve the accuracy of machine learning algorithms. High dimensional data can affect the accuracy of the predictive models, as it includes redundant and less important features (Y. Wu & Zhang, 2004). Thus, the intention of this feature engineering is to diminish the high dimensional feature space by removing redundant and least important features.

The Recursive Feature Elimination with Cross-Validation (RFECV) feature selection technique was chosen for building a feature space. Guyon et al. (Guyon, Weston, Barnhill, & Vapnik, 2002) proposed Recursive Feature Elimination (RFE) for gene selection and obtained an improved classifier performance. RFE is a widely used method for feature selection, which has been implemented in several studies (Shao, Yang, Gao, Zhou, & Lin, 2017; Youssef, Mohammed, Hamada, & Wafaa, 2019; Richhariya, Tanveer, Rashid, Initiative, et al., 2020; Alshabandar, Hussain, Keight, & Khan, 2020; I. M. K. Ho, Cheong, & Weldon, 2021).

During the process of Recursive Feature Elimination, feature importance is calculated in each iteration and the redundant and least important features are pruned from the current feature set until an optimal number of features is obtained. Each updated set of features was used to recalculate the performance of the model. The optimal number of features is identified by using cross validation with Recursive Feature Elimination, where a score is computed on the validation data. The features that give the maximum score on the validation data are selected as the optimal feature set.

5.2.3 Hyper Parameter Tuning

Every machine learning algorithm has hyper parameters that need to be configured before running the models on a dataset. Hyper parameter tuning is one important factor that needs to be meticulously configured to optimise performance of these machine learning models. Enhancing the performance of machine learning algorithms helps to outperform the existing benchmarks and sets new benchmark performances for a given problem/-domain (Snoek, Larochelle, & Adams, 2012). Additionally, different methodologies can be compared if they use identical model settings, thus hyper parameter tuning promotes fair comparisons. Furthermore, hyper parameter tuning helps improve reproducibility and diminishes the human effort required (Bergstra, Yamins, & Cox, 2013).

This study focuses on two different hyper parameter tunning techniques: Grid search and Random search (Liashchynskyi & Liashchynskyi, 2019). Grid search attempts to find the best possible combination from the given set of hyper parameters. It works by building a model for each combination of all the given parameters and selects the best model by evaluating them. Conversely, in Random search, a statistical distribution of hyper parameters is given and the values are randomly sampled to build the model.

5.3 Experiment

The dataset for this study was extracted from the AdelaideX[1] platform 'Introduction to Project Management' course offered in 2016. A total of 8,300 discussion forum posts were extracted from 982 different users. During pre-processing, 63 non-English discussion forum posts were removed before building the predictive models. The dataset was given to the annotators in three different iterations to annotate the forum post as mentioned in section 4.7. The detailed explanation of the data, including data collection, pre-processing and data annotation, is presented in Chapter 4. The descriptives of the corpus data is presented in Table 5.1 .

TABLE 5.1: Corpus Descriptives

Learner Role	Number of Posts
Information-Giver	4790
Information-Seeker	885
Other	2562

As mentioned previously, feature extraction is performed using the LIWC tool. According to feature extraction, 93 linguistic features were extracted for each learner post. To identify the optimal feature space, highly correlated features (> 0.8) were eliminated from the dataset using the Pearson Correlation Coefficient. These highly correlated features will contribute to the redundant feature space and do not improve the model performance drastically. Subsequently, during the feature engineering process, feature selection was performed using Recursive Feature Elimination with Cross-Validation, as mentioned in section 5.2.2 on the discussion forum posts, to rank the linguistic features. During the feature engineering, Random Forest Recursive Feature Elimination with cross-validation model was trained using 10-fold cross validation.

Classifiers like Random Forest, Logistic Regression, and Decision Tree were built from the scikit-learn library, while boosting algorithms like XGBoost were implemented using xgboost library. The classifiers were built using 10-fold cross validation. Prior to building these classifiers, to eliminate the class imbalance across the target class (information-giver, information-seeker and other), the class imbalance was handled using class weights.

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

Finally, the best performance model was evaluated on other discussion forum posts retrieved from the 'Risk Management for Projects' course to ensure the validity of the model. The Risk Management course contains 685 information-givers, 151 information-seekers and 361 other users.

5.4 Results

This section presents the results obtained from the role modelling process. It reports the classifier performance of different learning algorithms and also presents a statistical analysis of significant features that were obtained through feature importance. Finally, it reports the results of classifier performance from the model evaluation.

5.4.1 Learner Role Identification

Multi-class classifiers were constructed to detect learner role in a discussion forum post, addressing the first research question of this study: 'RQ1: To what degree of granularity can a machine learning model predict learner roles in discussion forum posts using linguistic features alone?'.

According to the feature ranking with RFECV, 42 optimal features were chosen as the best set of features to build the machine learning models. With these selected linguistic features, a multi-class classifier was built with several machine learning algorithms. Each model was constructed using 10-fold cross validation with weighted average evaluation metrics. Moreover, hyper-parameter tuning was conducted for each classifier to obtain its best performance.

The results of the classifier performance using 10-fold cross validation is reported in Table 5.2. It shows the classifier performance using several measures (i.e., precision, F-measure, and recall).

According to the hyper parameter tuning for the Random Forest classifier using the scikit-learn library (RandomizedSearchCV and GridSearchCV), the results show that the Random Forest classifier performs at its best in the following parameter setting:
criterion: 'entropy'
max_depth: 11

TABLE 5.2: Classifier Performance for Learner Role Identification

Classifier	Precision	Recall	F1-score
Logistic regression	0.86	0.84	0.84
XGBoost	0.87	0.87	0.87
Decision Tree	0.81	0.81	0.84
Random Forest	0.88	0.87	0.87
K-Nearest Neighbors	0.75	0.77	0.76
AdaBoost	0.85	0.85	0.85

max_features: 'auto'

min_samples_leaf: 4

min_samples_split: 2

n_estimators: 200

With an 87% F-measure, the study has demonstrated that analysing the language of post content itself is sufficient to predict 'information-seeker', 'information-giver' and 'other' roles in a discussion forum post. Therefore, it is evident that linguistic features have a high impact on learner role prediction in discussion forums.

5.4.2 Feature Importance

The study involved in calculating the feature importance of the linguistic features that are extracted from discussion forum data to address the second research question of this study: RQ2: What are the linguistic features that contribute significantly towards identifying a learner role that is demonstrated in a forum post?

Initially the feature importance algorithm calculates how each linguistic feature decreases the impurity of the split. Afterwards, the feature importance is calculated by averaging the decrease in the impurity over all the trees in the forest for a given feature. According to the feature importance, a 42 optimal feature subset was selected using the Random Forest-RFECV. The top 15 linguistic features that significantly contribute towards identifying a learner role are presented below:

1. Question Mark (QMark)

2. Words Per Sentence (WPS)
3. Word Count (WC)
4. Interrogatives (interrog)
5. Personal Pronouns (i)
6. Differentiation (differ)
7. Causal (cause)
8. Pronouns
9. Personal Pronouns (shehe)
10. Articles
11. Tone
12. Auxiliary verbs (auxverb)
13. Affect
14. Analytic
15. Cognitive Process (cogproc)

Subsequently, the Kruskal-Wallis statistical test is performed in the extracted 42 linguistic features to determine whether there are statistically significant differences across learner roles.

The results of the Kruskal-Wallis H Test for the top 15 features are presented in Table 5.3. This table summarises the results for the two main roles (i.e., information-givers, information-seekers) that were identified in this study.

TABLE 5.3: Results of Kruskal-Wallis H Test

Linguistic Feature	Learner Role	Mean Rank	Significance
Question Mark	Information-giver	3914.95	0.0E0
	Information-seeker	6677.32	

Words per Sentence	Information-giver Information-seeker	5270.73 3240.10	0.0E0
Word Count	Information-giver Information-seeker	5310.35 3097.48	0.0E0
Interrogatives	Information-giver Information-seeker	4614.96 5244.12	0.0E0
Personal Pronouns (i)	Information-giver Information-seeker	3573.51 4580.39	$2.6417E^{-144}$
Differentiation	Information-giver Information-seeker	4702.11 4755.97	$3.0023E^{-270}$
Causal	Information-giver Information-seeker	4912.16 4164.82	0.0E0
Pronouns	Information-giver Information-seeker	3466.15 5255.61	$1.4381E^{-190}$
Personal Pronouns (shehe)	Information-giver Information-seeker	4726.91 3393.42	0.0E0
Articles	Information-giver Information-seeker	4896.90 4199.73	0.0E0
Tone	Information-giver Information-seeker	3419.94 3891.19	$5.8523E^{-290}$
Auxiliary verbs	Information-giver	4164.49	$3.3836E^{-98}$

	Information-seeker	5507.35	
Affect	Information-giver	3569.62	$1.4633E^{-184}$
	Information-seeker	3828.97	
Analytic	Information-giver	4674.90	$3.595E^{-154}$
	Information-seeker	2728.38	
Cognitive Process	Information-giver	4519.53	$1.0682E^{-149}$
	Information-seeker	4870.76	

5.4.3 Model Evaluation

The best performance model has been evaluated with the discussion forum posts obtained from another MOOC to address the final research question of this study, 'RQ3: To what extent can machine learning models that rely on linguistic features be used across courses from similar domains?'

The classifier model was built on discussion forum data retrieved from the 'Introduction to Project Management' course and selected the Random Forest model as the best performing classifier. To ensure the validity of the model and check whether the model built using the project management data could perform better on another course, the model has been evaluated on the discussion forum data derived from another course that was similar in nature.

The 'Risk Management for Projects' course delivered by AdelaideX in 2017 was selected as the evaluation dataset. A total of 1197 learner posts were annotated, as mentioned in section 4.7. The same methodology was followed in this study to extract the features using the LIWC tool. The significant features that were identified during the 'Feature Engineering' (see 5.2.2) process for the 'Introduction to Project Management' discussion forum data were selected.

Figure 5.2 shows the precision, recall and F-measure obtained for the 'Risk Management for Projects' course. Due to the class imbalance, weighted evaluation measures

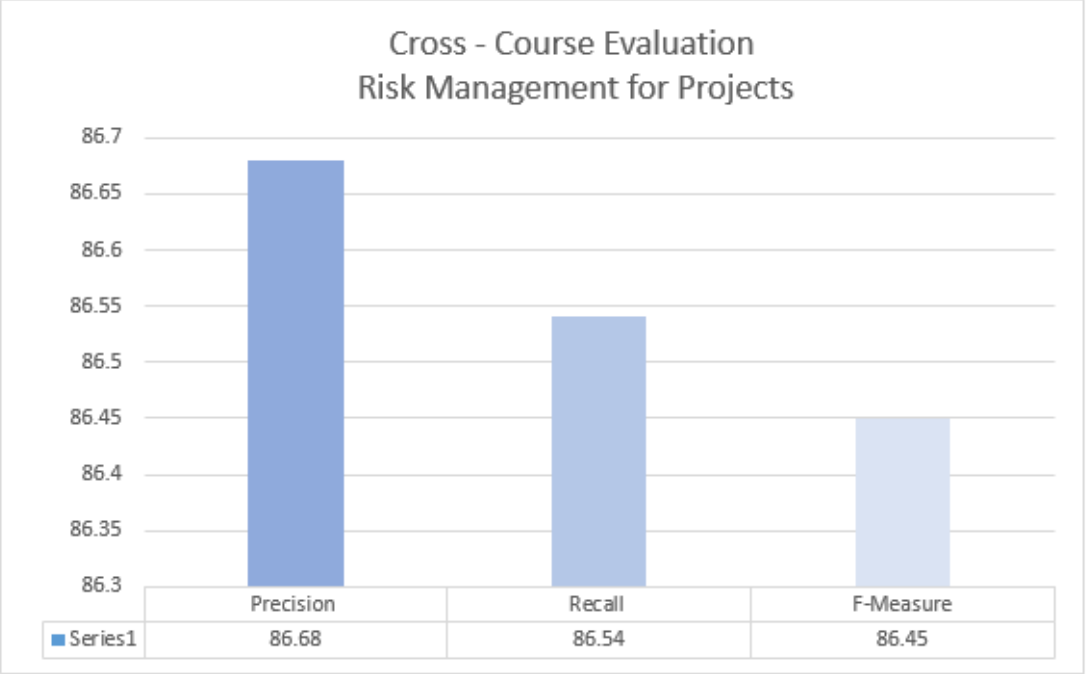


FIGURE 5.2: Cross-course Evaluation Results for Risk Management for Projects

were used, whereby the evaluation measures were calculated using the weights of each target variable. According to the results, the study obtained an 86% F-measure for the Risk Management course with 42 linguistic features. The study results confirm that it is possible to build a linguistic-only model for learner role identification in a forum post. Furthermore, the study proves that linguistic-only features are made effective in role identification by eliminating structural features such as thread positions and votes. This shows that real-time role identification is achievable without any delay to wait for structural features or network formations. Moreover, the model evaluation shows the identified linguistic features and the machine learning model developed in this study can be used in learner role identification in similar courses.

This study demonstrates that a machine learning model, built using the role modelling methodology of a previous semester's data, can be deployed in an on-going course to predict learner roles in discussion forum posts. Therefore, this methodology can be adopted by other courses (e.g., medicine and computer science) that are not similar to the domain used in this study.

5.5 Discussion

The end goal of this thesis is to investigate the student learning in Massive Open Online Courses. To investigate students' learning, this thesis chose to identify and examine the learner role and linguistic expression and built further investigations with these measures.

The prevailing literature emphasises the importance of learner collaboration in identifying user roles. In a computer supported collaborative learning environment, sufficient learner collaboration is vital to identify learner roles, as the majority of the studies focus on social network analysis. However, this study intends to identify the learner role in a discussion forum post to promote role prediction in forums where sufficient collaborations (i.e., sufficient replies for discussion forum threads) do not happen. A detailed explanation about the corpus used in this study is described in Chapter 4.

Additionally, most research studies (S. N. Kim et al., 2010; Bhatia et al., 2012; Arguello & Shaffer, 2015; Hecking et al., 2016) have conducted post classification in different domains by incorporating contextual features (e.g., structural features) separate from discourse features. However, the aim of this study is to identify a learner's role within a discussion forum post in a real-time MOOC environment without any delays in the predictions. Therefore, this study extended the line of research and built a supervised learning model to identify learner roles (information-seekers, information-givers and others) using real-time linguistic features extracted from their discussion forums posts. The following subsection discusses the key findings by addressing the research questions related to this study.

5.5.1 RQ1: To what degree of granularity can a machine learning model predict learner roles in discussion forum posts using linguistic features alone?

The existing literature (Strijbos & De Laat, 2010; Dowell & Poquet, 2021) has identified different user roles such as, 'Captain', 'Pillar' and 'Socially Detached' in discussion forums using network measures (e.g., weighted indegree, weighted outdegree) and group communication measures (e.g., participation, overall responsibility, social impact). Apart from the data limitations (i.e., less structured discussion forums), another important consideration is that these roles (e.g., Captain) can potentially be involved in both seeking

and giving information. For example, Strijbos and De Laat (2010) define 'Captain' in collaborative learning as 'A person who invests a lot of effort in the collaborative task'. Moreover the Captain ensures that the students work together as a group in the collaborative environment. The qualities of a 'Pillar' are a mirror image of the role 'Captain' but can be seen in larger group settings. There is a possibility, that either a 'Pillar' or a 'Captain' in a network can continuously seek information across different threads. Neglecting to identify the seeking behavior in a discussion forum post might lead to many drawbacks in real-word massive open online courses. Failure to identify a seeking behavior will lead to a confusion state in learning and eventually increase the dropout rate. Therefore, identifying seeking and giving behaviour of a learner is an important aspect in any online learning setting. Hence, this study identified the learner roles as information-givers, information-seekers and others in a given forum post.

In the existing literature (S. N. Kim et al., 2010; Bhatia et al., 2012; Arguello & Shaffer, 2015; Hecking et al., 2016), learner roles and post classifications in different domains have been predicted based on contextual features such as votes and views. These contextual features might delay the predictions in a real-time environment. In a real-time system, it is not realistic to wait for the context-related features (e.g., votes, views) for the post classification as they occur throughout the course and may change over time. Therefore, this study extends the line of research to construct a machine learning model to identify learner roles (e.g., Information-givers and Information-seekers) using linguistic features alone.

To address RQ1, an experiment was conducted on discussion forum posts to identify the learner roles. The success of this approach, with an 87% of F-measure, demonstrates that linguistic features can be used in real-time learner role predictions. This model can be applied in less structured MOOC environments where minimal collaboration can be observed. Furthermore, it is a domain-independent model, as it only considers discourse features.

According to the feature space used in this classification, it is evident that 42 linguistic features are significant enough to distinguish clearly information-givers, information-seekers and others. Furthermore, the results also demonstrates that the LIWC tool has enough potential to identify learners' roles in MOOC discussion forums posts.

5.5.2 RQ2: What are the linguistic features that contribute significantly towards identifying a learner role that is demonstrated in a forum post?

To address the second research question of this study, the Kruskal-Wallis H Test was performed on the linguistic features.

The results of the Kruskal-Wallis H Test demonstrate that all 42 linguistic features are significantly different between the information-giver and the information-seeker. Further analysis was conducted on the main two learner roles (i.e., information-givers and information-seekers) for the top 15 linguistic features that were identified during the feature engineering.

According to the feature importance, the 'Question Mark' is the most highly predictive feature in classifying learner roles. It is apparent that the question mark is a measure of information-seeking behaviour in any context. The mean rank of question marks between information-giver and information-seeker demonstrates (information-giver: 3914.95; information-seeker: 6677.32) more questions are being asked by information-seekers, which is their evident behaviour.

Why do you consider Risks Associated (or risk assessment of the project) only during the execution phase? If I start Planning and leave the risk assessment out of the scope, then during the execution phase I might be confronted with costly risks that can compromise the project and my responsibilities. What are your comments? **- Information-Seeker**

The second and third highly predictive features are 'Words Per Sentence (WPS)' and 'Word Count'. These features are an indication to identify how much a learner is writing (i.e., using more words) when reflecting their thoughts. According to the descriptive statistics, WPS is significantly high in information-givers with a mean rank of 5270.73; whereas information-seekers have a mean rank of 3240.10. According to literature (Arguello et al., 2006), high amount of words per sentence is an indication of linguistic complexity. This shows, students who are information-givers demonstrate high linguistic complexity in their discourse. Similarly, Kruskal-Wallis H showed that there was a statistically significant difference in word count between the learner roles with a

mean rank word count of 5310.35 for information-givers and 3097.48 for information-seekers. These results indicate that information-givers tend to use more words when sharing their thoughts than information-seekers in discussion forums.

In Anne's case, I think the budget is very important since it is determined by how much the couple wants to spend. Wedding events can easily cost thousands and thousands of dollars and the higher the budget, the more spectacular the event can be arranged. If the couple has a lower budget, Anne needs to carefully choose her resources to create a successful event. The tight deadline of 8 weeks makes it even more difficult, since she might face higher charges due to short notice bookings. **- Information Giver**

Similar to 'Question Marks', 'Interrogatives' are question-like function words (e.g., how, when, what, who and where), which are also significantly high amongst information-seekers with a mean rank of 5244.12 compared with information-givers. This shows that information-seekers use more question-like clauses than information-givers.

I have a question. In the initiation phase, how did you deal with the concept of "options"? What did you consider in that? **- Information-Seeker**

"I'm a bit uncomfortable with the face to face meeting system because of its likely downsides; majorly proximity between rendezvous and the stakeholders. What if they are all far apart? How do you manage that?" **- Information-Seeker**

It can be observed that the overall use of pronouns is high, with information-seekers having a mean rank of 5255.62 compared with information-givers. A similar trend is noticed with first person singular pronouns ('I'), where a 4580.40 mean rank is held by information-seekers. This suggests that information-seekers self-reflect in their posts by relating content to themselves. A study by Atapattu et al. (2019) identifies personal pronouns ('I') are higher among confused learners. This shows information-seekers who incorporate more personal pronouns ('I') in their discourse, express their confused state of mind.

I was confused with this concept also. I've chosen the Anne's projects and I thought that her "options" were to ask the couple if she can make any change without their

supervision. I've read in others posts that "options" is the different ways the project manager has to completes his tasks, if that's the case I think we must try to think what we would do in such cases and create different ways to success in the project. I'm still waiting for answers. **- Information-Seeker**

Hi everyone, I also have a question about this task. Maybe you can help. I can do the exercise and reorder the boxes. But I can only have them all start at the same day. Which means they would be implemented simultaneously. This can't be correct. Can you also change the start date of each task or is the exercise really just changing the order? Would be happy to receive feedback on this. Thanks a lot! **- Information-Seeker**

Causation is one of the sub elements of the cognitive process explaining the cause and effect of an event. According to the analysis, the causation is high in information-givers with a mean rank of 4912.16 compared with information-seekers. Conversely, the results show that the cognitive process including differentiation is significantly high in information-seekers compared with information-givers.

What method is used to determine the flow rate and tight design ?. I consult this because in a similar case I use two methods, dominant flow and energy balance and this produced two very different degrees of security situations, which influences the cost / resources of the project. **- Information-Seeker**

Articles, which are a type of determiner, are significantly higher in information-givers, with a 4896.90 mean rank, than information-seekers. The study by Gundel et al. (1993) states greater use of determiners is a reflection of more giveness in a text, which indicates text cohesion.

LIWC defines Analytical thinking in two different dimensions namely logical hierarchical thinking and personal narrative thinking. The analysis shows that information-givers hold a mean rank of 4674.90 demonstrating the logical hierarchical thinking of learners whereas information-seekers posses 2728.38 mean rank.

Hi. I analyzed Anne's project as well. I didn't assess the supplies as a high risk, once the venue was established. I thought the short lead time would allow them to make

choices in the context of what was available. I like your strategy to transfer the risk of the dress to the bride. That seems more appropriate. Did you consider using the PESTLE framework in analyzing this project? I found it helpful in getting me to think about things that I would not normally. It helped me to really walk through the project and think about all things that might go wrong; not just the ones that readily came to my mind. **- Information-Giver**

Moreover, the analysis confirmed that the emotional tone expressed by information-seekers is statistically significantly high with a mean rank of 3891.19 compared with information-givers. Similarly, the overall affective process, which considers both positive and negative emotions, is also high in information-seekers with a mean rank of 3828.97.

According to the initiation template i feel confused between the project objectives the deliverable of the project . could any of my teachers or course mates help me how to distinguish between them ? **- Information-Seeker**

Through this language analysis, the study has demonstrated the characteristics of the learner role using the highly predictive linguistic features. These linguistic indicators show great potential to understand a users' behaviour in any kind of discussion forum. As linguistic analysis is a separate line of research, it opens the doors for many researchers to generate different linguistic behaviours associated with the user role in any discussion forum.

5.5.3 RQ3: To what extent can machine learning models that rely on linguistic features be used across courses from similar domains?

In regards to the third research question (RQ3), related to cross-course evaluation, the model was evaluated to validate the predictive model built using only the discourse features. Model evaluation is another important aspect of this study, to ensure the re-usability of the model. Cross-course evaluation is considered to encompass building a model on one single course and then evaluating the model on data from another course.

During the evaluation, the model achieved good performance with an 86% F-measure for unseen data retrieved from the 'Risk Management for Projects' course. Moreover, the

results confirm that 42 linguistic features are sufficient for the learner role identification when performed on discussion forum data from a similar course domain.

Furthermore, the results also validate that the methodology used above to build the predictive model can be re-used in another entirely different course where this predictive model cannot be used for predictions. For example, a medicine course or software engineering course does not fall under the domain used in this research study; therefore, a practical implication of this proposed methodology would be that building a machine learning model on a previous semester's data and implementing the model on the ongoing course would still predict real-time roles in discussion forum posts.

5.6 Summary

This chapter presents the first study of this dissertation aimed at identifying the learner role in discussion forum posts. The study identifies the learner roles as information-giver, information-seeker and other using the grounded theory approach. This chapter describes the detailed methodology of role modelling, including feature extraction with the LIWC tool, feature selection using RFECV and hyper parameter tuning with Grid search and Random search. This methodology can be replicated in other MOOC courses to identify learner roles that best describe their forum data.

With an 87% F-measure, this study proves that a machine learning model with linguistic-only features is capable of predicting learner roles in discussion forum posts. The model can identify learner roles in real-time, as it only uses the discourse features. Furthermore, the model was evaluated on a similar course extracted from the AdelaideX learning platform. The evaluation results ensure the re-usability of the role identification model.

Additionally, this study presents the top linguistic features, according to feature importance, that contributed significantly towards learner role identification. According to the statistical analysis, linguistic features like question marks, interrogatives, and personal pronouns are high in the information-seeker's discourse. Conversely, information-givers use more words per sentence, have a higher word count and show greater analytical skills in their discourse. The statistical analysis confirms features that were obtained through the feature selection are significantly different across the different learner roles. These

significant features that distinguish the learner roles can be used as language indicators in a learner's discourse.

The next chapter of this dissertation presents the study on topic modelling in learner posts and visualises the relationship among these topics and learner clusters, including the learner role identified in this chapter.

Chapter 6

Topic Modelling

6.1 Introduction

The primary goal of any learning medium is to enrich the knowledge of learners regardless of how much it can contribute towards learners' knowledge acquisition. In an learning environment, individual learners reflect different levels of knowledge improvement. A learners' knowledge acquisition can be improved by activities such as answering learners' queries, knowledge-sharing and explanations. With thousands of participants in a MOOC learning context, it is really challenging to identify the degree of each learners' understanding of the delivered learning content.

Identifying the lecture topics or key terms that are understood and not understood by the learners is important. Exploring the relationships across lecture materials (i.e., course topics), learner roles (e.g., information-giver, information-seeker) and different learner clusters (i.e., High Distinction information-givers, Fail-grade information-givers) has the potential to develop further understanding of student learning and students' levels of understanding on a given lecture topic. Identifying such relationships between learners and lecture materials helps instructors to identify the lecture topics that were not understood by the students. Such information can guide course leaders as to where extra attention should be given during the course period and improvements and redesigns can be enforced for subsequent course offerings.

To have a better way of understanding a large corpus like MOOC discussion forums, tools and techniques are required to perform automatic analysis. Topic modelling is

one such technique that supports the disclosure of the hidden structures prevailing in a corpus. Topic models are probabilistic generative models that have been widely used to automatically organise, search and summarise a large electronic corpus. Topic Modelling can be defined as a task that automatically identifies topics from a given set of documents (Blei, 2012). Given a document, it aims to uncover the most relevant and appropriate topics that best describe it.

This thesis aims to use topic modelling on learners' discussion forum posts to identify the topic being discussed across different learner clusters (e.g., information-givers versus information-seekers and High Distinction learners - Fail-grade learners). By exploring such relationships within different clusters, it can help to understand student learning patterns as the course progresses. Moreover, it also helps instructors to come up with learning strategies and course designs that will both increase retention and support learners' knowledge achievements.

In order to address the aforementioned aim, this study addresses the following research questions:

RQ1- What are the main discussion topics discussed in learner posts?

Learners who are enrolled in MOOCs can post a wide range of information in discussion forum posts in which they can seek or give information. Identifying major discussion topics in discussion forums is important as they can help course providers and learners in many ways. Firstly, identifying issues or problematic lecture topics raised by learners will help instructors and their peers with instant knowledge-sharing. This will also help course designers to redesign certain course units that had issues, or raised questions in previous course offerings. Moreover, this will also help learners and instructors to easily find the solutions provided for questions or issues raised during the course. Discovering such solutions will help instructors to have quick validations to ensure correct knowledge has been shared among peers.

RQ2 - What are the main discussion topics discussed in different learner clusters?

In a real-world scenario, each learner does not gain an equal amount of knowledge during the learning process. This results in different learner clusters, such as information-givers, information-seekers, High Distinction information-givers and Fail-grade information-seekers.

These different learner clusters can post questions or provide answers in different topic areas. It is important to classify such discussion topics relevant to different learner clusters as they can provide useful insights. For instance, identifying information-seeking behaviour from low grade learners is important to deliver prompt instructor interventions that can help such learners to progress further. This will also help to reduce the drop-out rates, and number of confused learners, whilst supporting a reduction in misconceptions. Moreover, knowledge-sharing (i.e., information-giving) posts from low grade learners need additional monitoring from instructors to avoid any misleading information being given to the learning community. Hence, there is a high chance of making such online –learning mediums like a traditional learning environment, in which a personalised learning can be given for different learning clusters. Thus, analysing learner topics across different learner clusters will provide many benefits.

It is expected that addressing the above research questions will provide a way to investigate learning by means such as identifying the course topics that were frequently used to seek information or topics that were dominated by Fail-grade learners. Furthermore, the findings of this study will help the learning community to create strategies that can be implemented during the course to increase the retention rate.

This chapter is summarised as follows: Section 6.2 describes the methodology used to predict discussion forum topics, including the steps involved in training an LDA topic model on an extended training corpus followed by topic extraction. Section 6.3 presents the results of the topic model and describes the results by addressing each research question of this study. Section 6.4 details the discussion on the topic model along with the major contribution and findings of this study. It also presents the strategies that can be implemented in a MOOC learning environment. Finally, Section 6.5 summarises the chapter.

6.2 Methodology

This study used the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic model to extract topics from lecture materials (i.e., the transcript from lecture videos embedded in the course) and learner posts. LDA is a statistical generative model that can be used to discover hidden topics in documents as well as the words associated with each topic.

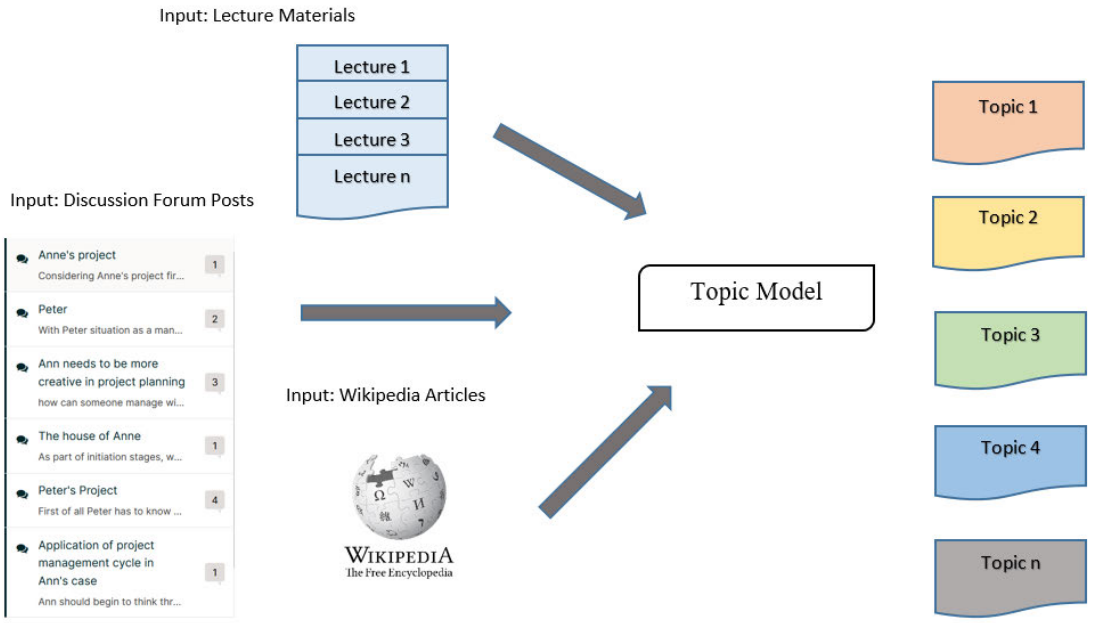


FIGURE 6.1: Overview of the topic modelling methodology

Figure 6.1 describes the overview of the topic modelling methodology used in this study. This study trained a topic model using the LDA algorithm on an extended training corpus (i.e., lecture materials, Wikipedia, learner posts). Lecture transcripts of the 'Introduction to Project Management' course were identified as the initial training corpus. This course has a total of six learning modules covered during the semester that include:

- WEEK 1: What is Project Management?
- WEEK 2: The Initiation Phase
- WEEK 3: Understanding the Planning Phase
- WEEK 4: Determining Project Risk
- WEEK 5: Project Teams and Communication
- WEEK 6: Bringing it all together

For each week, an average of nine sub-lectures were presented, and each of these sub lectures were treated as a document. It is vital to enhance the semantic space to correctly identify topics from learner posts. To ensure that the semantic space majorly reflects the topics being discussed in the forum posts, the training corpus was extended using

the seeding method (Cai et al., 2018). This ensures suitable contextual information for key terms obtained from the learner posts is given during training. Figure 6.2, presents the methodology used to extend the training corpus.

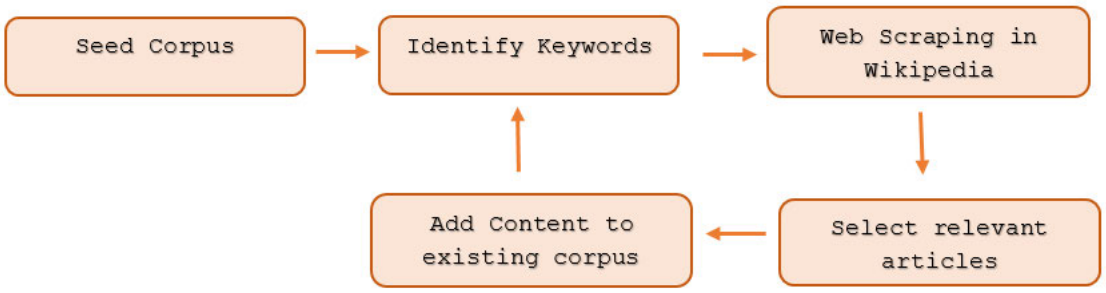


FIGURE 6.2: Methodology to extend a training corpus

To this end, discussion forum posts from the 'Introduction to Project Management' course delivered during a 'different semester' were added to the original corpus to enrich the training corpus. Lecture transcripts and these discussion forum posts were identified as a seed corpus. A seed corpus can be defined as a number of documents that represent the domain. To determine the most frequently discussed words and keywords, frequencies of words were analysed from these lecture transcripts and forum posts. N-grams (sequences of words of length N) that were extracted from the lecture transcripts are presented in Figure 6.3 and Figure 6.4.

To start with, a meticulous analysis is performed on the n-grams to identify relevant n-grams from both the lecture transcripts and the discussion forum posts for web scraping. Table 6.1 shows the relevant n-grams used for the Wikipedia search.

Furthermore, word clouds were also used to visualise the lecture materials obtained from the 'Introduction to Project Management' course, to look into the major key terms that occurred frequently in these lecture materials. Similarly, these word clouds were also used to visualise the discussion forum that was used in the training corpus. Word clouds have become a popular visualisation technique that serves as a very first method to understand the overview of text documents (Burch, Lohmann, Pompe, & Weiskopf, 2013; Sinclair & Cardew-Hall, 2008). They are used widely in various contexts, particularly to analyse data derived from online platforms. Word clouds are also integrated with several text analysis systems such as OpinionSeer to analyse customer reviews (Y. Wu et al., 2010) and Jigsaw for investigative analysis (Stasko, Görg, & Liu, 2008). A word cloud is a

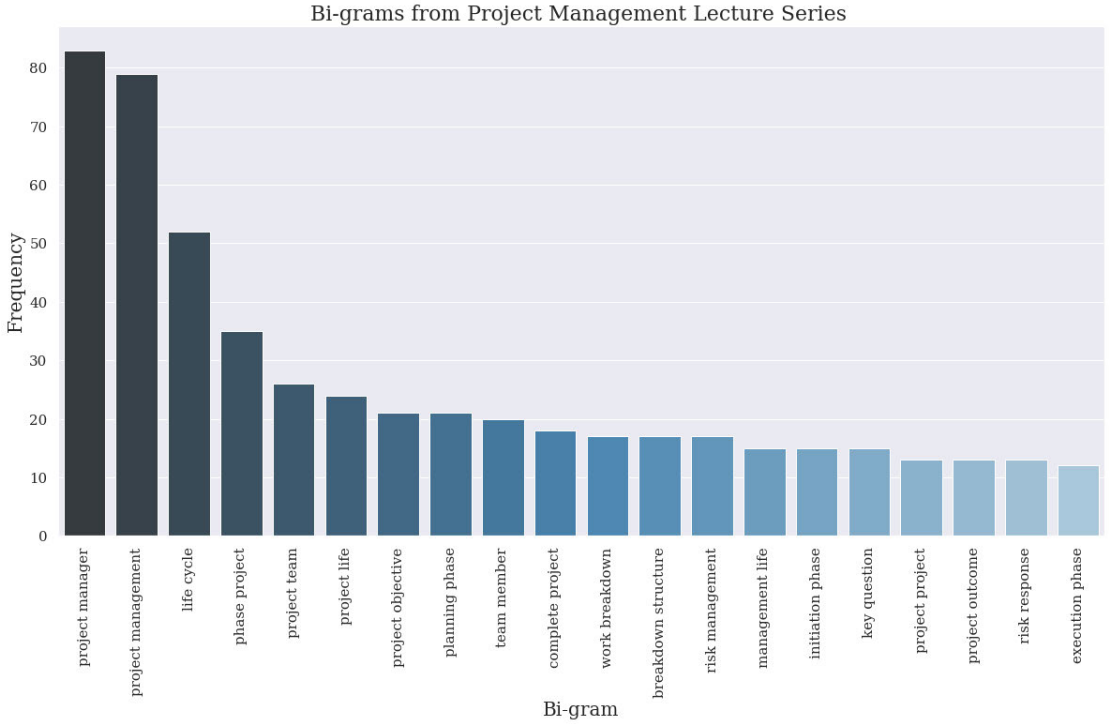


FIGURE 6.3: Bi-grams extracted from the lecture materials

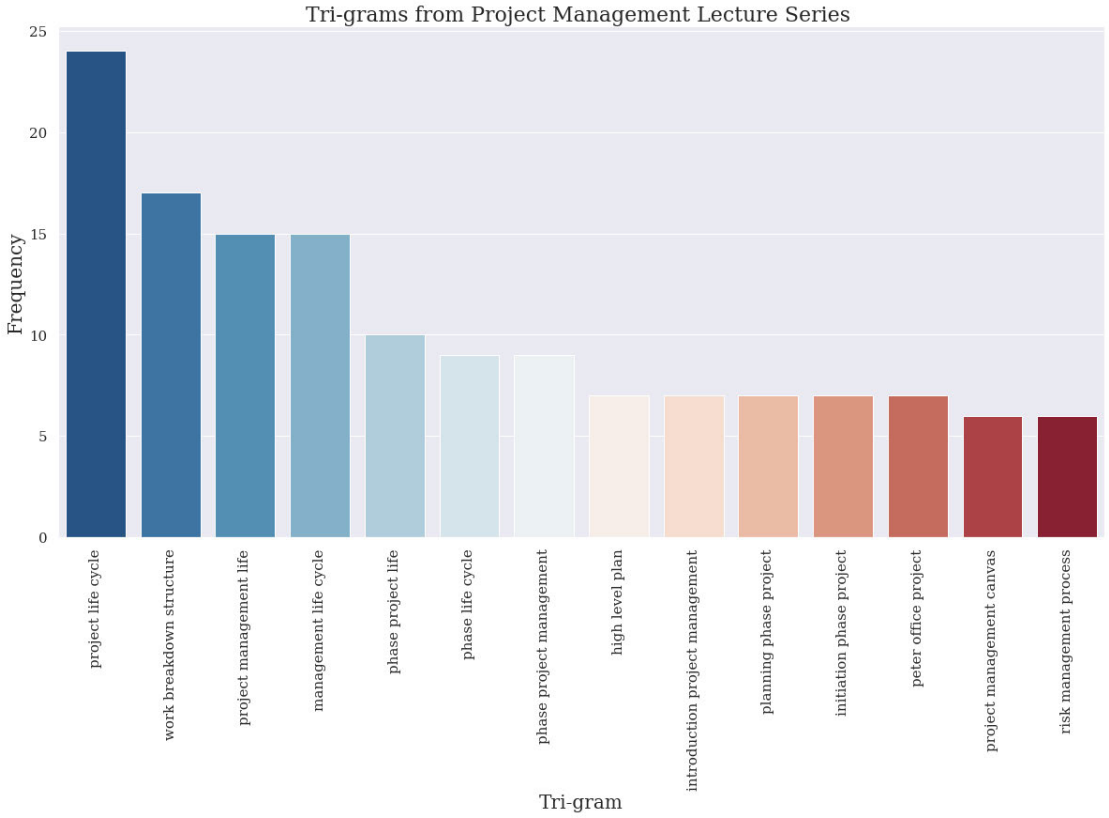


FIGURE 6.4: Tri-grams extracted from the lecture materials

simple, yet nevertheless powerful, visualisation method where each word is shown in various sizes indicating its frequency or importance in a given corpus. In other words, it visualises the most frequent word with large and bold letters and less important words in a smaller size (Lohmann, Ziegler, & Tetzlaff, 2009). Therefore, this study used these word clouds to visually track the key terms in the training corpus (e.g., lecture materials and discussion forum posts).

TABLE 6.1: N-grams used for Wikipedia Search

Material	N-grams
Project management lecture series	Stakeholder Project Risk Project management Project manager Initiation phase Planning phase Project team Risk management Risk response Project life cycle Risk management process
Project management discussion forum posts	Stakeholder Project Initiation phase Project management Project manager Planning phase Project management skill

With these keywords, web scraping was performed on Wikipedia, a free online repository. It has been proven that using Wikipedia for external knowledge can enhance the performance of document clustering (X. Hu, Zhang, Lu, Park, & Zhou, 2009). Therefore, to further extend the document representation, these newly identified documents from Wikipedia are then added to the training corpus and keyword extraction is repeated to identify new keywords from the extended corpus. This process was repeated until the

final corpus rationally represented the dataset (discussion forum posts) in this study. Such that topics that are deviated from the discussion forum posts were not included in the final corpus.

After identifying significant words from lecture transcripts and discussion forum posts, a Wikipedia API[1] was used to extract the content from the wiki pages. Moreover, Wikipedia 'search' was used to retrieve the most searched queries relevant to the given word. For example, the search result for 'Project Management' returns the title of the following wiki pages, as shown in Figure 6.5.

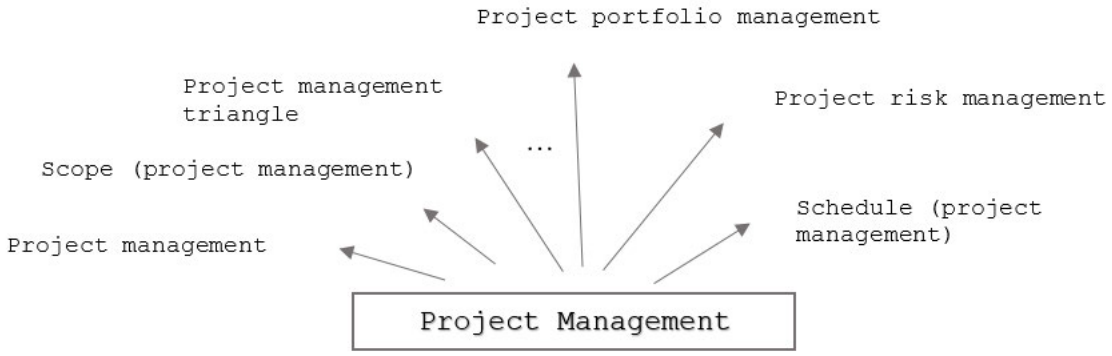


FIGURE 6.5: Wikipedia search result for "Project Management"

As aforementioned, wiki pages were retrieved for each significant term obtained from lecture materials and discussion forum posts. These significant terms were obtained through analysing the n-grams. Moreover, the relevant wiki pages of the search queries were also included in the training corpus. Nevertheless, the final corpus was constructed only from those relevant wiki pages that highly resemble the topics/ significant terms discussed in the lecture materials and discussion forums. Wikipedia pages were manually examined to ensure its relevance to the lecture transcripts. The selected wiki pages were included entirely in the corpus. The descriptive information for the initial corpus and the extended corpus is presented in Table 6.2.

[1]https://pypi.org/project/wikipedia/

TABLE 6.2: Descriptive Statistics of the Project Management Corpus

Dataset	Unique Terms
Initial corpus with lecture transcripts	1612
Expanded corpus with Wikipedia and additional forums posts	27982

A useful topic model always requires meticulous data cleaning. Therefore, the most common natural language pre-processing steps were performed on the training corpus, as presented below:

1. Convert the text to lowercase.
2. Remove numbers and punctuation.
3. Remove stop words - Words such as 'for', 'is' and 'of' are the most common words that add no significant meaning to the text. These words were removed.
4. Lemmatisation - Lemmatising words reduces words to their common base form (i.e., their roots). These words can be treated as different entities by the topic model if they are not lemmatised. This will eventually reduce the significance of these words in the topic model.

After pre-processing using the aforementioned first three steps, phrase modelling was performed prior to Lemmatisation. Phrase modelling is a technique to learn meaningful frequent multi-word concepts in a corpus that are derived from different token combinations. In other words, it identifies frequently used phrases rather than identifying every n-gram in the given corpus. The study identified phrases with the words that co-occur frequently by looping over words in the training corpus. This helps to reduce the memory requirements by reducing the vocabulary size. This study used the underlying concept by Mikolov et al. (2013) to detect phrases. The following formula is used to determine whether two words A and B form a phrase.

$$score(A, B) = \frac{count(AB) - count_{min}}{count(A) \times count(B)} \times N \tag{6.1}$$

- count(A) - Number of times token A appears in the corpus
- count(B) - Number of times token B appears in the corpus

- count(AB) - Number of times token A B co-occur in the corpus
- N - The total size of the corpus vocabulary
- $\text{count}_{\text{min}}$ - A user-defined parameter to ensure that accepted phrases occur a minimum number of times.

The study identified a range of threshold (e.g. 1 to 20) and iterated over the threshold range to improve the accuracy of the topic model. Moreover, the study also identified the number of bi-grams under each threshold range. After considering these aspects, a threshold value was determined for phrase modeling. Bigrams with a score above a determined threshold are selected as phrases as follows: a threshold value of ten was selected to determine the degree of relationship between two tokens to accept them as a phrase.

$$\text{score(A,B)} > \text{threshold}$$

As LDA can be constructed with n-grams (e.g., uni-gram, bi-gram, tri-gram), an LDA model with bi-grams along with unigrams was developed. It has been proven that the quality of text analysis and text categorisation can be enhanced when using bigrams in addition to unigrams (Tan, Wang, & Lee, 2002). A sample of uni-grams and bi-grams obtained from the training corpus are presented below:

"project", "management", "project_management"

Once the text is converted to tokens, a dictionary and document term matrix were created to run the LDA algorithm. The MALLET (Machine Learning for Language Toolkit) (McCallum, 2002) was used to build LDA models as it includes an efficient and scalable implementation of the Gibbs sampling algorithm when compared with the original LDA, which uses an online variational Bayes algorithm (Hoffman, Bach, & Blei, 2010). Furthermore, MALLET results in a high coherence score and the outcomes obtained from MALLET topic models are precise. Apart from LDA MALLET, LDA Multicore was also implemented in this study. The results revealed LDA MALLET generates more meaningful, human interpretable topics/ keywords/terms, demonstrating that it is more appropriate for interpreting topics compared with the original LDA topic model.

It is important to choose an ideal number of topics to identify efficient and human interpretable topics. An ideal number of topics can be selected by examining the topic coherence score. The score is a measure of how well a topic can be coherent (i.e., offer semantic interpretability) to a human (D. Newman, Lau, Grieser, & Baldwin, 2010). If a topic is coherent, the words that represent a topic will be semantically related. Topic coherence is calculated by considering the semantic similarity among high probability words in a topic (Stevens, Kegelmeyer, Andrzejewski, & Buttler, 2012).

The LDA topic model was constructed to calculate the topic coherence score for different numbers of topics with a range of 2 to 15. A wide range like 2 to 100 was not selected as the entire course spans six weeks, with six major topics, and to obtain more meaningful topics. Moreover, it has been observed that increasing the number of topics results in poor quality topics (Mimno, Wallach, Talley, Leenders, & McCallum, 2011). The study chose a topic model with 13 topics as it obtained the highest coherence score. Moreover, to ensure the validity of the topic model, the topic coherence score is calculated several times by re-running the LDA model to confirm that the highest topic coherence is obtained for the same number of topics.

To assign a meaningful label for each topic, a manual analysis was performed on the top 20 terms along with the weights of each term. Furthermore, the model was executed with the training corpus (e.g., lecture materials) to confirm its integrity with the descriptive topic labels. As aforementioned, this LDA topic model was trained on the extended corpus (i.e., lecture transcript,another semesters' discussion forum posts and Wikipedia pages). This trained topic model was saved and reused for predicting the topics of each learner post of the current study data for further evaluation.

6.3 Results

This section describes the results obtained from the topic model and examines the topics against different learner clusters. The results are presented as a way to understand students' learning in relation to the topics covered in the course. Major topics that reflect the study dataset are presented in section 6.3.1 followed by section 6.3.2, which analyses the learner topics in relation to different learner clusters.

6.3.1 Topic Extraction

This section addresses the first research question of the study: 'RQ1- What are the main discussion topics discussed in learner posts?'

As aforementioned in the methodology, n-grams (e.g., unigrams and bigrams) were extracted from lecture transcripts and discussion forum posts of the same course delivered during another semester to extend the training corpus of the topic model.

Figure 6.6 shows the frequency of words obtained from lecture materials using word clouds. As can be seen in the image, the most frequent words are related to the key concepts in project management such as project, risk, team, planning and stakeholder. The extracted word clouds were used as a means to validate the n-grams obtained from the training corpus.

FIGURE 6.6: Word cloud of lecture transcripts for the entire course

Subsequently a Wikipedia search was performed from the derived n-grams to further extend the corpus. Afterwards, pre-processing along with phrase modelling was performed on the extended corpus before building the topic model. To obtain the optimal number of topics that best represent the training corpus, topic coherence scores were calculated for a range of 2-15 topics. According to topic coherence, the highest coherence score was obtained for 13 topics.

The trained LDA model was applied to the corpus (e.g., lecture transcripts) and identified the topics discussed in each lecture material and forum post. To come up with a topic

label, the researcher manually reviewed the top 20 terms and analysed the most relevant content within each topic. After manually analysing the relevant content (i.e., content of lecture transcripts and forum posts) under each topic, two topics were eliminated as they primarily represented non-content posts such as those relating to peer-networking and introductions. Since the aim of this study is to investigate the student learning, the non-content posts were eliminated. Only the 11 remaining content-related topics were used for further analysis. A few examples of the non-content posts obtained from the topic modelling include:

"I just completed the course successful. Hope to get the certificate soon"

"Hi, <name>. Greetings from Taiwan! I'm just starting in this course and do not work in project management either. Hope this course will be useful to you as well while you ponder your career change. Best wishes!"

"Hi guys, let me introduce by myself, my name is <name>, I am from Monterrey Mexico and I am interested in this course because i want to apply all the tools in the day to day and in my work."

Table 6.3 presents the top 20 terms obtained for each topic from the topic model. Furthermore, the study applied word clouds for each topic to visually identify the significant words used in each topic. These word clouds helped the researcher to easily validate the meaningful descriptive label given for each topic. Figure 6.7 shows the word clouds obtained for each topic where the term weights are used to differentiate the importance of each word.

TABLE 6.3: LDA topics and its word representation

Topic	Topic Label	Terms per Topic
1	Project Planning & Triple Constraints	project, cost, schedule, scope, resource, budget, activity, task, janet, time, important, complete, determine, peter, define, time_frame, stakeholder, planning, scenario, plan

2	Project Communication	communication, tool, work, meeting, team, group, document, information, share, business, good, people, communicate, create, team_member, company, training, access, easy, person
3	Real-world Project Experiences	time, start, day, give, end, make, decide, set, goal, thing, place, finish, organize, people, work, month, find, prepare, task, long
4	Janet's Case Study	time, work, material, tile, find, complete, project, job, move, budget, house, cost, buy, floor, money, hire, order, janet, kitchen, build
5	Anne's Case Study	anne, wedding, couple, event, plan, week, budget, venue, option, date, bride_groom, problem, thing, idea, location, party, happy, vendor, detail, guest
6	Evaluating and Monitoring Project Progress and Project Handover	project, process, management, company, team, design, customer, department, development, software, manager, system, datum, issue, requirement, include, develop, site, implementation, support
7	Project Lifecycle	plan, project, client, planning, execution, phase, time, require, initiation, set, order, involve, resource, identify, closure, deliver, understand, stage, clear, expectation
8	Project Stakeholder and Project Team Management	project, team, involve, stakeholder, manage, manager, objective, lead, question, work, outcome, goal, involved, engage, research, success, achieve, leader, information, role

9	Peter's Case study	peter, system, project, problem, community, stakeholder, young_people, option, government, develop, school, job, month, understand, create, people, community_organization, employment, provider, organization
10	Project Risks	risk, change, area, project, high, due, case, issue, social, technology, deal, delay, involve, low, impact, avoid, legal, reduce, supplier, factor
11	Real-world Project Experiences	challenge, year, program, product, work, service, benefit, provide, date, design, write, month, receive, goal, end, staff, company, include, period, start_end

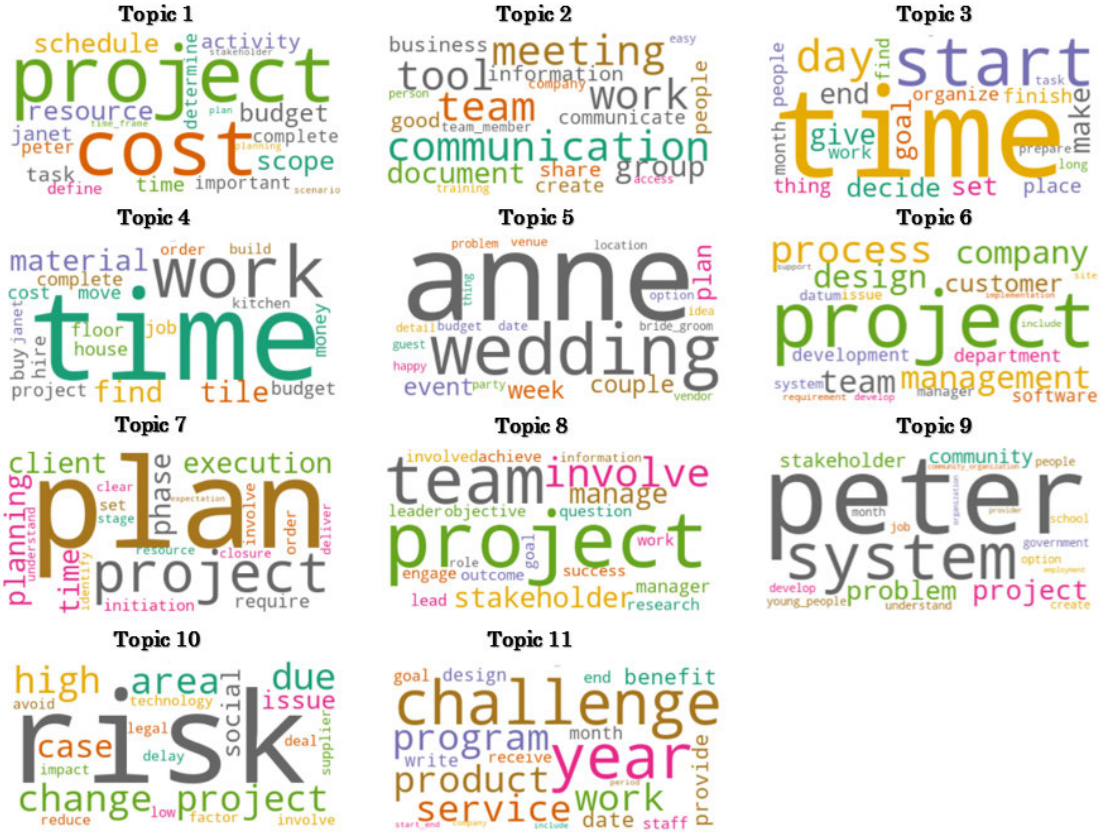


FIGURE 6.7: Word clouds for topics extracted from LDA

Next, the trained LDA model was applied on the learner posts of this current research study to determine the topics being discussed in each discussion forum post. Figure 6.8

represents the topic distribution of this study data. According to the statistics, 'Project Communication' is the topmost topic that has been discussed by learners while 'Janet's case study' has been discussed least.

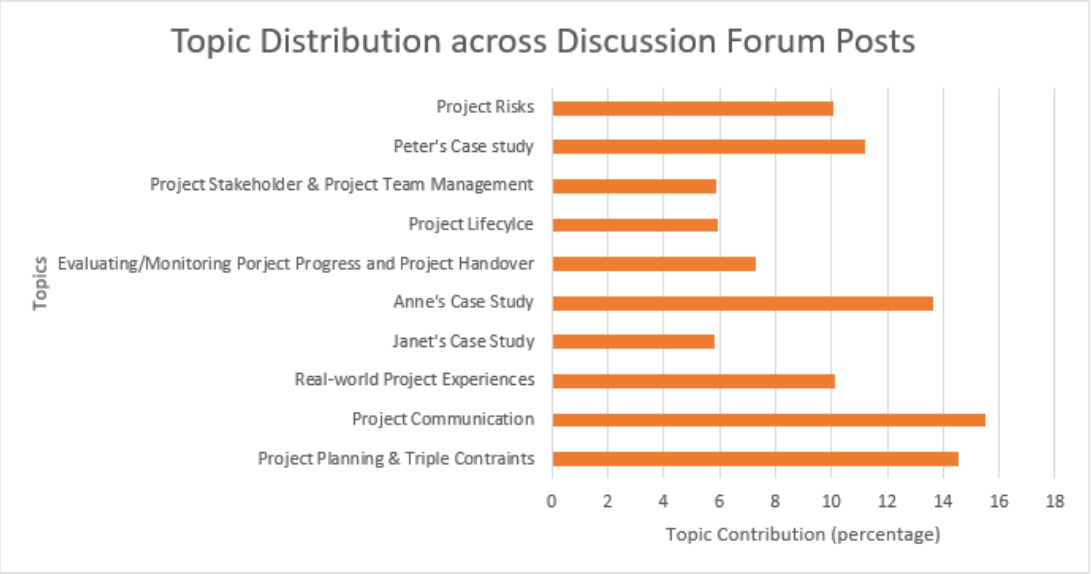


FIGURE 6.8: Topic distribution across discussion forum posts

Subsequently, these identified learner topics were visualised against different learner clusters as seen in the next section.

6.3.2 Topic Distribution with Learner Clusters

This section addresses the second research question of this study:'RQ2: What are the main discussion topics discussed in different learner clusters?'. In this study, the identified learner clusters are defined as a given role by type (information-giver, information-seeker) and by role according to grade attainment (e.g., High Distinction information-giver, Fail-grade information-giver, High Distinction information-seeker, Fail-grade information-seeker).

6.3.2.1 Topic Distribution across Learner Roles

After identifying the topics of learner posts, the study classified these learner posts along with their relevant topic into learner roles (i.e., information-giver and information-seeker) as identified in Chapter 5 - Role Modelling.

Figure 6.9, shows the topic distribution between information-givers and information-seekers. Since topic 3 and topic 11 (see table 6.3) discuss 'Real-world Project Experiences', the study merged these two topics as a single topic for the analysis.

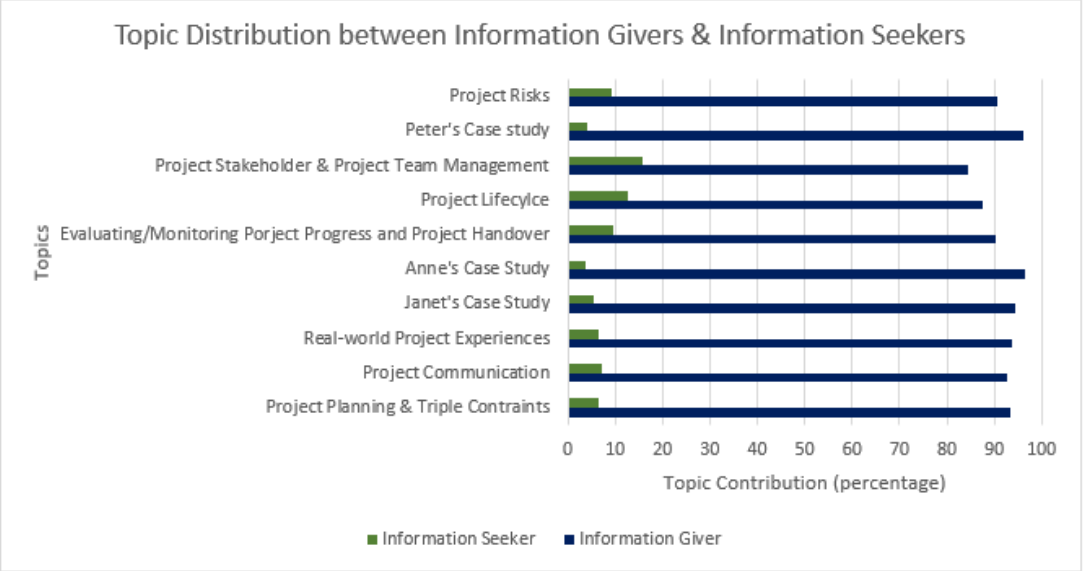


FIGURE 6.9: Topic distribution across learner roles

According to Figure 6.9, a higher topic contribution is given by information-givers than information-seekers throughout the course. The course structure for the dataset involves a question posted by an instructor every week, inviting learners to contribute to the discussion. Such instructor directions may be the major reason for the learner's active contributions as 'information-givers' in the discussion forums.

A few sample questions posted by an instructor in the discussion forum are presented below. Detailed information about the structure of the course is presented in section 3.3.1.

Week 1 –

In this exercise, we would like you to provide an example of a memorable project that you were involved in:

- *Describe what the project was;*
- *Describe the challenges you faced, in particular related to the (4) key elements of all projects.*

Week 5 –

In this exercise we want you to share your favourite communication tools and describe how you use them effectively. If you aren't familiar with this type of technology, then start with those mentioned in the tools and strategies video and investigate further from there.

This study divided the overall course structure into three sections for further evaluation. Figure 6.10 presents the course division followed in this study.

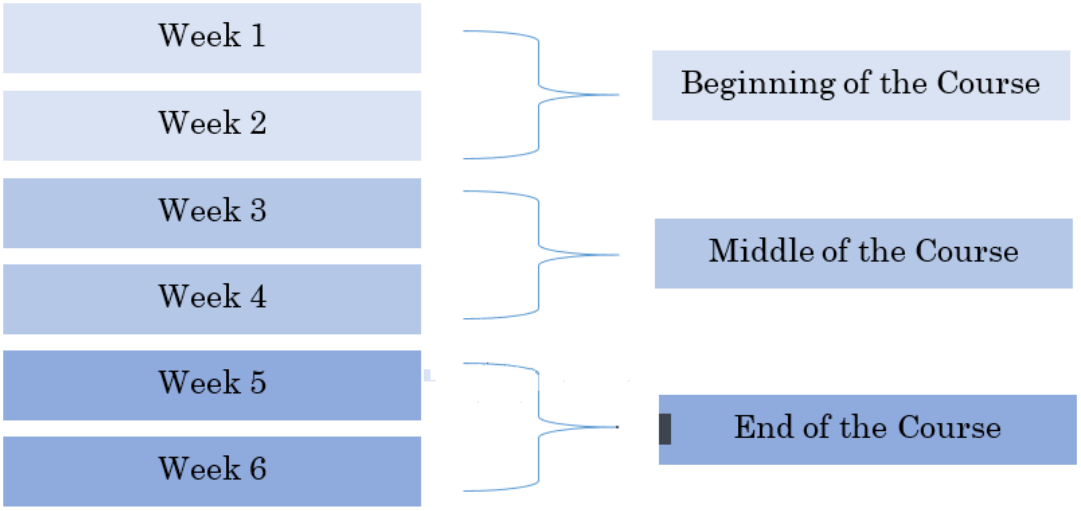


FIGURE 6.10: Course Structure

According to Figure 6.11, the top five major contributions of the information-givers can be seen in 'Project Communication', 'Project Planning & Triple Constraints', 'Anne's Case Study', 'Peter's Case Study' and 'Real-world Project Experiences'.

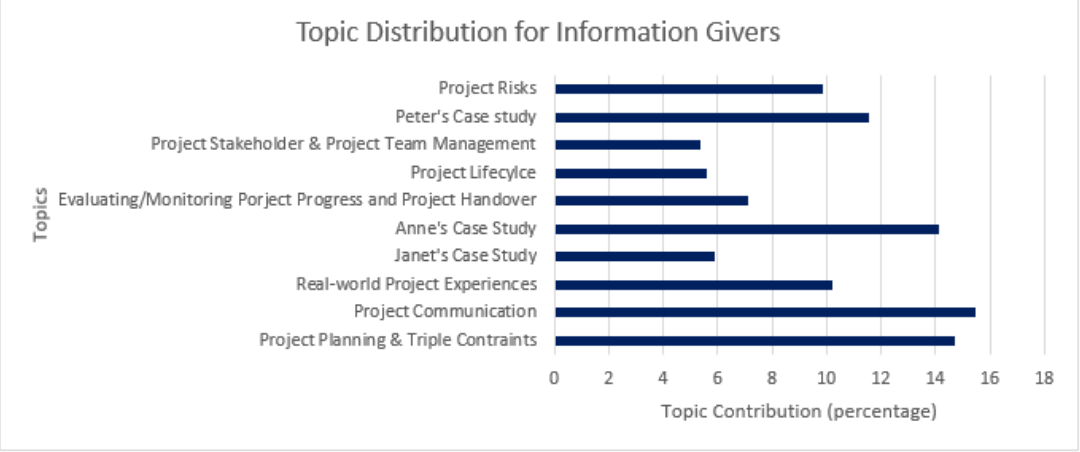


FIGURE 6.11: Topic distribution for information-givers

According to the course structure, 'Project Planning & Triple Constraints' relates to the content covered in week 3 (i.e., the middle of the course). The week 3 question posted by the instructor expects to apply the concepts related with project planning to the case-studies that were presented in the course. Since learners can easily engage with the case studies, high contributions were observed for this module. Furthermore, this module resulted in high information-seekers, thus answering the information-seeking queries might result in high information-giving posts.

"1. Office Project - I think the most important piece of planning for Peter's Office project will be the measurable tracking of the implementation and progress of his plan. 2. DIY Project - Janet's most important planning item will be her budget, with little experience in remodeling she will need to be able to plan for unforeseen problems. 3. Wedding Planning - Anne's challenge in her project will be her schedule, she will need to carefully scope and track each of her deliverable to keep from missing her completion." **- Project Planning & Triple Constraints**

According to the lecture materials, 'Anne's Case Study' and 'Peter's Case Study' represent case studies that illustrate real-world project scenarios to reflect certain concepts in project management. Apart from Janet's case study, the case studies of Peter and Anne obtained high information-giving posts, revealing that introducing case studies will enable the students to engage more in discussion forums. Case studies open up opportunities for learners to give suggestions, opinions and share their knowledge and understanding related to these real-world examples. Examples of posts derived from Peter's and Anne's case study are as follows:

"Since Peter has little experience working with young people and community organisations, he also needs to enlist the help of people familiar with those two groups to move the project forward." **- Peter's case study**

"In this case, I believe the main aspect is time management since she only has 8 weeks to put together the wedding. Money is always a concern, but it seems that the bride and groom are more concern about to have it on the date they set for their wedding." **- Anne's Case study**

Similarly, analysing the topic 'Real-world Project Experience' reveals that these information-giving posts are related to the very first discussion forum question (Week 1) posted by the instructor. These contributions show that learners try to engage more in the discussion forums during the beginning of the course. Similarly, asking learners to share their own experiences that reflect a course's content will enable them to share their knowledge and engage more in that given forum thread.

An interesting pattern can be observed in the 'Project Communication' module that was delivered at the end of the course. This module contains high information-giving and information-seeking discussion forum posts, which reveals learner engagement increases towards end of the course. Another interesting interpretation can be made that a high percentage of information-seekers resulted in a high percentage of information-givers, as they try to answer the information-seeking learner posts.

"There are indeed limitations to cloud technology. I work with really big files and even with good connections, the uploading/downloading times can be really long and you can run out of storage space easily. The second limitation is privacy, I work in Canada,and, according to local laws, patient information cannot be stored in servers outside of Canada, so you cannot work with Dropbox in that case. " **- Project communication**

"Hi, there are two variants of SharePoint, one is SharePoint online which is cloud based where data is stored in Microsoft's servers, the other is share point server where the server resides in our organization but managed by Microsoft. " **- Project communication**

Conversely, as presented in Figure 6.12, major information-seeking behaviours are observed in 'Project Communication', 'Project Risk', 'Project Planning & Triple Constraints' and 'Project Stakeholder & Project Team Management'.

According to the lecture materials, 'Project Communication' represents the topic covered towards the course ending. Similarly, 'Project Stakeholder & Project Team Management' relates to the major topic covered at Week 5. This shows that the majority of information-seekers try to ask more questions on the modules that were delivered towards the end of course.

FIGURE 6.12: Topic distribution for information-seekers

"I find that for work. stopping by someone's desk is a great way to grab attention, especially after sending them a couple of emails that go unanswered and maybe calling them and they either promise to work on your issue and don't or they never answer nor call back! What are some of the limitations you find with Google Apps?" **- Project Communication**

Apart from the end of the course lectures, a high percentage of questions have been raised on 'Project Risk', which is delivered in the middle of the course. Analysing the discussion forum posts of 'Project Risk' shows that learners are interested in seeking information on the information-giving posts that were related to case studies and real-world examples on Project Risk, as they could easily engage with the real-world scenarios.

Further, analysing 'Project Risk' shows that the entire week was dedicated to this single topic with 8 kinds of lecture materials. This shows that focusing on a single topic throughout a week, can provoke learners' information-seeking behaviour, as they were curious to apply/see the in-depth concept in the real-world examples.

"By looking to outside consultation and now adding an additional financial cost to the project, are you increasing the risk in other "economic" areas? How would this affect the overall risk of the project? Increasing one of the PESTLE risk numbers to decrease another." **- Project Risk**

The course module 'Project Planning & Triple Constraints' has the highest number of learning activities (4) allocated in the entire course, including 3 knowledge checks along with 11 types of lecture materials. Even though this course topic has been delivered in the middle of the course, the module has a high contribution rate given by information-seekers because high peaks in information-seeking can be seen whenever there is an upcoming knowledge check (assignment). With more sub-topics delivered under a certain module and with additional learning activities, it is possible to observe an information-seeking pattern in discussion forums.

"Estimating the resources required and the duration is done during Scheduling. Why is it done again during Costing? Isn't it duplicating matters " **- Project Planning & Triple Constraints**

"as the project manager can optimize the activities and costs?" **- Project Planning & Triple Constraints**

Overall, this investigation reveals many takeaways with respect to learner roles. Firstly, learners engage more towards the end of the course irrespective of learner roles. Analysing the discussion forum posts along with their learner roles discloses that instructor directions and course structure influence the roles that learners undertake in forum participation. A question posted by an instructor in a discussion forum thread will have a high impact on whether a learner is going to be an information-giver or -seeker. For example:

"In this exercise we want you to share your favourite communication tools and describe how you use them effectively."

The aforementioned instructor post will lead the learners to be information-givers rather than information-seekers. Furthermore, posting a question as follows will likely lead the learners to become information-seekers.

"In this exercise we would like to share your queries or issues on a communication tool"

Moreover, course structures like weekly assignments and learning activities provoke the information-seeking behaviour of learners. Conversely, posting case studies in the lectures

will enable the learners to become information-givers as they try to give opinions and share their knowledge.

The study further divided the posts into 'Comment Thread' and 'Comments' to understand the learners' behaviours in the different levels of discussion. Figure 6.13 shows the topic distribution identified at different discussion levels.

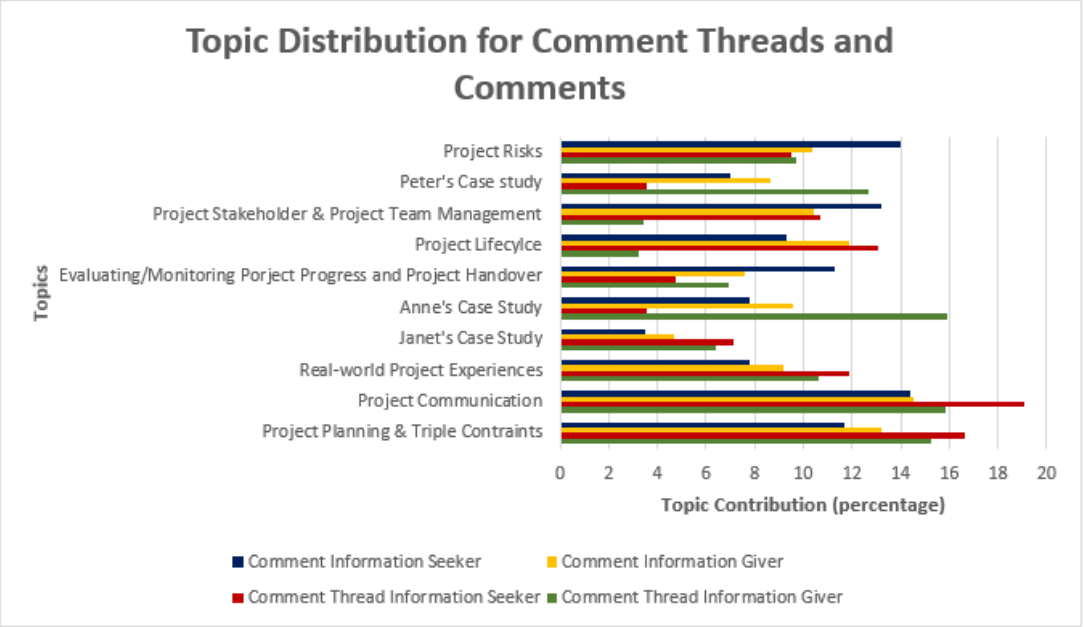


FIGURE 6.13: Topic distribution identified in Comment Threads and Comments

Even though an instructor asked learners to share their knowledge on a certain topic, learners initiated the discussion forum to seek information on topics such as 'Project Communication', 'Project Planning & Triple Constraints' and 'Project Lifecycle'. This shows the learners' lack of understanding of the given lecture content. Identifying such topics will also help the lecturers and instructors to consider learning strategies that might progress the learners' understanding in these course topics. These strategies can be considered at two levels, as follows:

1. Course improvement in the following semester
2. Immediate improvement in the on-going course

The first strategy can be addressed by improving or redesigning the identified lecture topics from the previous semester in a logical or comprehensible way. It can be implemented by identifying the topics that had the highest number of information-seekers and

modifying the learning content based on the most frequently asked questions under each identified topic.

Nonetheless, it is hard to develop instant strategies for an on-going course but strategies for improving the on-going course content can be derived from Figure 6.12. For example, elaborating topics via case studies will help learners to easily understand the given topic as fewer information-seekers were observed. Therefore, instructors might introduce brief case studies during the course to explain the topics that were not understood by learners. Furthermore, instructors can also provide additional reading materials for the topics that were less well understood.

Moreover, instructors can see the topics that may need interventions by visualising the information-seekers against the course topic. From Figure 6.12, the 'Project Communication' and 'Project Risks' topics required immediate interventions. Furthermore, from Figure 6.13, 'Project Planning & Triple Constraints' can also be identified as needing interventions as it occupied the second highest information-seeking position in the Comment Thread. Therefore, implementing these strategies in a MOOC environment helps learners to understand the course content more effectively by providing additional resources or prompt instructor interventions.

Since the data collected for this study reflects a sample of all the discussion forum posts and did not capture the entire body of comments that falls under a single Comment Thread, analysing the relationship between Comment Threads and Comments can give misleading information. Thus, as future work, extracting all the comments under each Comment Thread could reveal interesting observations. For example, it is possible to calculate the contribution of information-giving comments given for a particular information-seeking forum thread. This shows how much an information-seeking question is being addressed by learners' comments, or whether there are any further information-seeking comments being observed. Nevertheless, with the topic modelling and role modelling methodologies, it is possible to further extend this study.

6.3.2.2 Topic Distribution with Course Grades

This study also conducted an experiment to explore the relationship between student grades and the topics that were identified using the topic model. As the course delivered

by AdelaideX, the study used the grading scheme that has been widely used by The University of Adelaide, South Australia. Thus, the study used the grading scheme as mentioned in the University's policies[2]. The detailed grading scheme is mentioned in Table 7.1 (Grading Scheme from a traditional learning environment) and is also attached in the Appendix (see Appendix A).

Figure 6.14 shows the topic distribution among different grade distributions for information-givers. The results show information-givers who obtain High Distinctions always contribute in any discussion forum thread, irrespective of the topic. On the other hand, interesting patterns can be observed in most of the topics, such as information-givers who obtain Fail grades participate more than the information-givers who obtain Pass and Credit in the final grading on course topics such as 'Peter's Case Study', 'Anne's Case Study' and 'Project Stakeholder & Project Team Management'. Moreover, Fail-grade learners participate more than Distinction grade learners on certain topics such as 'Project Lifecycle', 'Project Stakeholder & Project Team Management'.

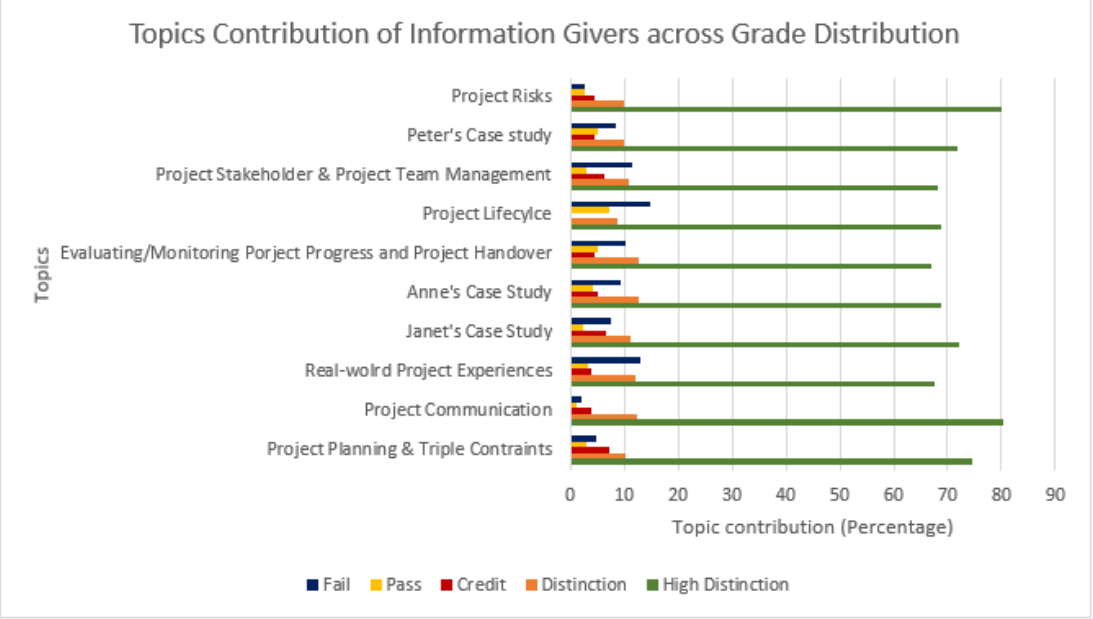


FIGURE 6.14: : Topics contribution of information-givers across grade distribution

Figure 6.15 presents the topic distribution among different grade distributions for information-seekers. According to Figure 6.15, it shows a similar trend observed for information-givers across different grade distributions. The results show information-seekers who obtain a High Distinction more frequently seek information across topics than other groupings.

[2]https://www.adelaide.edu.au/policies/700

In most of the topics, learners who obtain Fail grades seek information more than the learners who obtain a Distinction in final grading on course topics such as 'Project Risks' and 'Project Lifecycle'. Similar trends can be observed between Fail and Pass, Credit learner grades across the topics such as 'Peter's Case Study' and 'Anne's Case Study'.

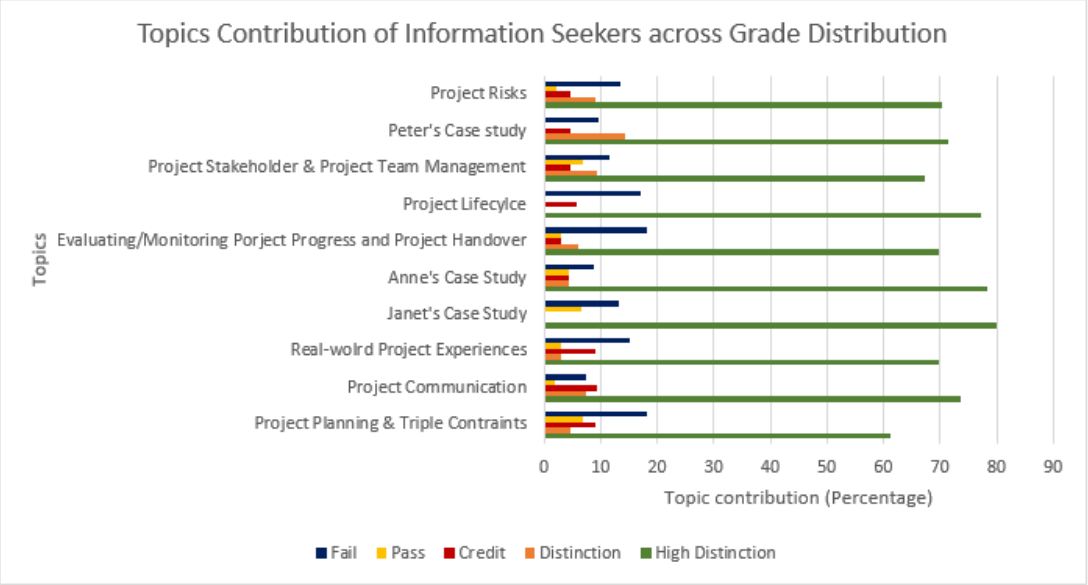


FIGURE 6.15: : Topics contribution of information-seekers across grade distribution

Analysing the student roles relevant to student performance shows that learners' engagement patterns in discussion forums do not always have a positive correlation with learner achievements. For instance, active learners who always post on discussion forums cannot be necessarily considered as high performance students (e.g., comparing Distinction and Fail grades |Pass and Fail grades). There can be various reasons that can force such participation, such as learner engagement may be a mandatory requirement in a course. Moreover, high confusion can also lead to such discussion forum participation.

According to the study by Wong et al. (2015), active users are considered as users whose participation in discussion forums is active and constant during the course. The results derived from Figure 6.14 and Figure 6.15 show that overall Fail-grade students are actively engaging in discussions more than students who obtained Distinction, Credit and Pass grades, irrespective of their learner roles. In other words, active users cannot always be considered as high performers (i.e., bound to obtain good grades. e.g., High Distinction or Distinction); therefore it is important for instructors to take further steps to validate these posts. Lack of monitoring and overdue feedback can facilitate misleading information to other learners.

A study by Wong et al. (2015) also investigated 'do active users generally make a positive contribution to the MOOC forum?'. One of their findings show that active users have a higher proportion of posts with negative votes than typical users. Interestingly, in this study, Fail-grade students highly engage in discussion forums more than Distinction, Credit and Pass grade students. It is worth investigating the quality of the contributions of these discussion forum posts. One way to validate the content of the posts is to measure if these posts positively contribute to knowledge achievement. Other than manually reading the posts, a quick step to validate the content is by making use of other structural elements provided by discussion forums. For example, as suggested by Wong et al. (2015), votes (i.e., up votes, down votes) can be used to ensure whether or not a post contributes positively to learners' knowledge achievement.

These results show that identifying learner roles in discussion forums helps the instructors to tailor their feedback to specific student groups. For example, in this course, it is clear that Fail-grade information-givers tend to give more information than relatively high performance students (Distinction, Credit and Pass). Thus, the instructors should always verify their answers for the most crucial or significant topics that have been discussed in the course. This helps to avoid misleading answers and misconceptions occurring in discussion forums.

Moreover, the results also demonstrates that the second highest percentage of information-seekers are Fail grade students. Therefore, it is vital for instructors to provide timely feedback whenever they find information-seeking posts. This increases the chances of students who are at risk of failing obtaining the timely support that they need to pass the course.

The aforementioned discussion proves that identifying these learner roles helps instructors to tailor their feedback accordingly. For instance, if instructors identify that a particular group of students are always seeking information in discussion forums, rather than obtaining a High Distinction, there is a high probability that these learners' would obtain a Fail-grade towards the end of the course. Thus, instructors could redesign a particular section/module of the course that has a large amount of information-seekers in order to support these learners.

6.4 Discussion

This study is intended to identify the topics that were expressed in the discussion forum post, along with visualising them against different learner clusters to deepen understanding of students' learning. The major contribution of this study is the topic modelling methodology that was conducted to build the LDA topic model. This methodology explains how the topic model was trained on the extended corpus, which consists of lecture transcripts, and a discussion forum from another semester under the same course, alongside Wikipedia content. This pre-trained LDA model can be utilised and evaluated by other researchers to identify the topics on a similar course's discussion forum posts.

The next contribution of this study was the key takeaways that were derived by analysing these topics against different learner clusters. Analysing the topics against user roles demonstrated the ways in which instructors designed the weekly questions may be the foremost reason behind having more information-givers than information-seekers for each topic. Furthermore, the analysis also shows a high number of information-givers can be seen when introducing case studies to explain certain course concepts. Additionally, learners also actively emerge as information-givers while discussing their real-world project experiences. This shows that, real-world examples have a high impact in encouraging learners to think and share their knowledge.

Conversely, learners emerge as information-seekers during topics that were delivered especially at the end of course, notably topics such as 'Project Communication' and 'Project Stakeholder & Project Team Management'. The findings confirm that introducing more learning activities, such as knowledge checks, and delivering a module that consists of a wide range of topics (i.e., lecture materials) as well as delivering in-depth content, will also increase the chance of developing a wider range of information-seekers. The topics 'Project Communication' and 'Project Planning & Triple Constraints' have been identified as having highest contributions across both information-givers and information-seekers. It might be possible that learners become information-givers as they try to answer the questions posted by information-seekers. Therefore, in the future, analysing such relationships along with time aspects and forum elements (i.e., comment thread, comment) might reveal useful insights about the learning process.

The results of this study can be used in many ways. Firstly, the findings of this study can assist in course design to increase learner participation in forums. Secondly, several predictions, such as high information-seeking course components/units, learner roles (i.e., information-giver/ information-seeker) that might significantly contribute to different discussion topics can be derived from this analysis. According to these predictions, redesigning a course or constant instructor interventions over a given period of time (i.e., high information-seekers were seen at end-of the course), can be implemented to improve learners' overall learning.

The analysis conducted between topics and course grading schemes demonstrates that High Distinction learners make high levels of contributions across course topics irrespective of learner roles. The analysis shows that information-givers with Fail-grades share their knowledge more frequently than information-givers who obtain Pass or Credit as their final course grade. Likewise, Fail-grade information-seekers make a higher percentage of responses in most of the topics compared with the Distinction, Pass, and Credit grade learners. This analysis shows that learner engagements do not always positively correlate with learner achievements. According to the results, prompt interventions for information–seeking posts are required, as the majority arise from Fail-grade learners. Furthermore, information-giving posts with high negative votes need to be validated by instructors to prevent any misleading information being given during peer interactions.

The analysis discloses that it is not only important to identify topics but also investigate them across different learner clusters to help instructors to easily identify learners' levels of understanding of lecture topics. Such analysis will help the educators to generate learning strategies that will provide a more effective learning environment.

6.5 Summary

This chapter presented the topic modelling that was conducted on the discussion forum posts to better understand students' learning. The aim of this study is to discover the topics being discussed across several learner clusters such as information-givers versus information-seekers, Fail-grade learners versus High Distinction-grade learners and so forth. The methodology of this chapter outlined a detailed approach to building an LDA topic model on an extended training corpus to successfully predict the topics that

reside in learners' discussion forum posts. The topic model identified ten course-related topics that were later provided with meaningful names for further analysis. This topic model can be used as the training model to predict learner topics for subsequent course offerings.

Subsequently, this chapter presented a series of analyses that were conducted between topics and different learner clusters. The reasons for high information-seeking and information-giving course modules were presented. Furthermore, the findings confirm Fail-grade learner's actively participate as information-givers and information-seekers compared with Distinction, Pass and Credit grade learners in different topics; therefore, it is important to monitor their posts to avoid misleading information being given by them to their peers, and to minimise dropouts, as they often seek information. The study also presented a set of strategies that can be implemented in MOOCs to provide learners with a more effective learning environment.

The next chapter presents the study on the linguistic expressions that were extracted from the discussion forums. This study extracts rule sets from a predictive model that helps to predict the likelihood of course-grades based on the learners' linguistic behaviours. The chapter also visualises those linguistic features against course duration and finally presents linguistic profiles for two different learner groups.

Chapter 7

Linguistic Analysis

7.1 Introduction

The previous chapters investigated learner role identification and the relationship between learner clusters and their topic contributions in discussion forums. The investigation into the relationship between the learner clusters and their discourse topics revealed diverse correlations. Going beyond learner roles, this study aims to investigate the linguistic behaviours expressed in discussion forums to understand their contributions to identifying different learner groups, such as those who pass and fail.

Student success in courses has been explored in various studies under different names such as performance prediction (Okubo et al., 2017; Ashenafi et al., 2015; Dowell et al., 2015), predicting completion (Jiang et al., 2014; S. Crossley et al., 2015), predicting students at-risk (Aljohani et al., 2019; Marbouti et al., 2016) and dropout predictions (Sharkey & Sanders, 2014; Kloft et al., 2014). To understand students' learning, it is important to investigate more than performance prediction. Therefore, the main objective of this study is to understand the relationship between learner grades and the linguistic features that are extracted from learner discourse. In doing so, this study derives human understandable rule sets to explain the relationships. In order to achieve this, several machine learning models were built to understand the predictive capability of linguistic expressions, for predicting learner grades. At the outset, the study investigates grade predictions with different grade settings and also expands the feature space with time aspects to improve the model accuracy.

The next part of this chapter discusses the investigation conducted into the rule sets that have been extracted from the decision tree algorithm. These rules were built upon the linguistic feature space derived from the learners' discourse. It is anticipated that extracting a significant set of rules for investigation will help the learning analytics research community and educators to understand the relationship between the linguistic features and learners' grades specifically. This will reveal the hallmarks of specific linguistic behaviours by students who will achieve different grades in discussion forums, in terms of optional participation. The later part of this chapter visualises the contribution of linguistic features with time in two different components of the discussion forum (i.e., the Comment Thread and Comments) and defines a 'linguistic profile' for Pass and Fail grade learners using a set of linguistic features that differ significantly between these two learner groups.

To achieve the aim, the study is guided by the following research questions:

- RQ1: How do linguistic features extracted from students' discussion forum posts contribute to learner grade predictions?
- RQ2: What are the significant rules that can be developed using the linguistic features extracted from discussion forums to identify the likelihood of different learner grades?
- RQ3: What are the significant linguistic features that can contribute to developing linguistic profiles of learners?

Addressing these research questions will help to identify, distinguish learner behaviours of students who obtain different grades (i.e., Pass, and Fail), which can help instructors to intervene and prevent at-risk students from course attrition.

This chapter is structured as follows: Section 7.2 provides a detailed explanation of the methodology, including an explanation about extracting decision rules using entropy. Section 7.3 presents the results of the prediction model on different grade settings and presents an explanation of the decision rules. Section 7.4 provides a discussion on rule sets. Next Section 7.5 visualises the linguistic features across the course period. Section 7.6 presents the Linguistic Profiles of two different learning groups, followed by a discussion in Section 7.7. Finally, the summary of this chapter is provided in Section 7.8.

7.2 Methodology

This section presents the methodology of the study conducted to predict learner grades and rule extraction. An overview of the research methodology is provided in Figure 7.1, followed by a detailed description of the data and model building.

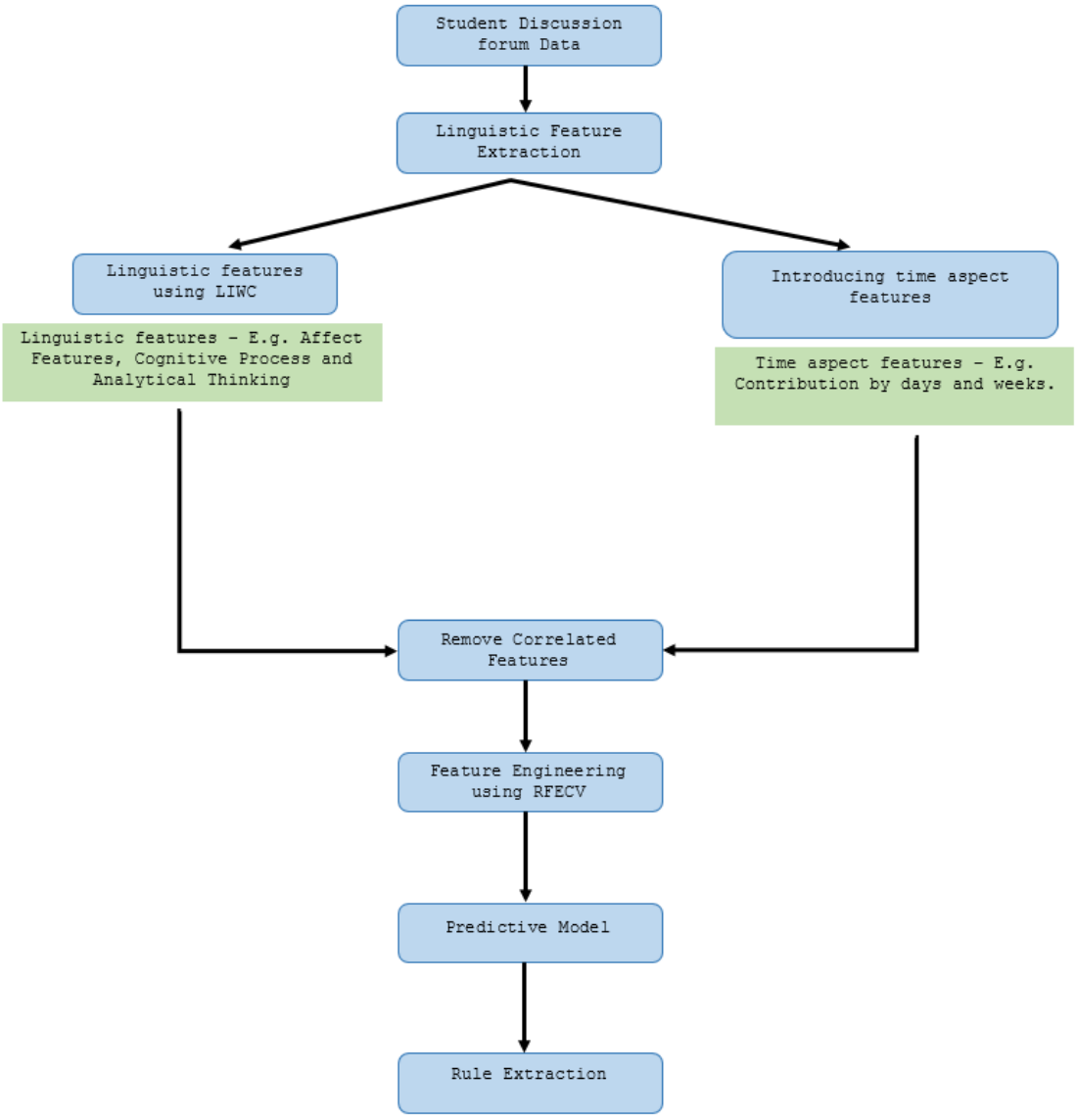


FIGURE 7.1: Overview of the methodology

7.2.1 Learner Grade Prediction

The dataset used for this study was collated from the 'Introduction to Project Management' course offered in 2016 on the AdelaideX[1] learning platform. This course has a dedicated discussion forum where learners can exchange knowledge and opinions in response to questions proposed by the lecturers and their peers. To conduct this experiment, entire discussion forum posts were extracted from the course. This resulted in a total of 26,605 user posts. The final data contained 22,300 posts from 6745 learners, after removing non-English posts and those posts from users who did not obtain a final grade. A detailed explanation of the research data, including the course format and learning context, is presented in Chapter 4.

The main objective of this study is to present human understandable rule sets to the learning analytics research community that can have practical implications. To extract those rules that could be further analysed to understand the relationship between linguistic expressions and learner grades, a series of predictive models were built for grade prediction to identify the significant linguistic features.

To begin with, the following pre-processing steps were conducted prior to the implementation of the machine learning models.

1. Remove non-English posts - The data includes 167 non-English posts. These posts were removed before proceeding further.

2. Grade Setting - Grades were assigned according to the guidelines built in Chapter 6. Refer to Table 7.1 for grade setting.

3. Feature Extraction - As aforementioned, discussion forums are the data source of this research study. To extract the linguistic features that are exhibited by learners, the study used the Linguistic Inquiry and Word Count (LIWC) tool designed by Pennebaker et al. (J. W. Pennebaker et al., 2001). The overview of the LIWC tool and its features, including the psychological aspect, are presented in Chapter 2 under Section 2.4.

4. Remove the correlated features - The features that are highly correlated (>0.8) were removed using the Pearson Correlation Coefficient. Having highly correlated

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

features does not improve the model's performance drastically, as they do not convey any extra information to the predictive model.

5. Feature Engineering - After building the feature space, feature engineering is performed using the Recursive Feature Elimination with Cross-Validation (RFECV) with 10-fold cross validation to identify the significant feature space that best contributes to learner grade prediction. A detailed explanation of feature engineering is given in Chapter 5 - Section 5.2.2.

The study initially categorised the learner grades into two different grading schemes, namely a grading scheme used in a traditional learning environment (i.e., grading scheme widely used by The University of Adelaide, South Australia) and the course grading scheme used in the given course. The detailed grading scheme used in the analysis is presented in Table 7.1.

TABLE 7.1: Grading Scheme

Grading Scheme	Outcome	Grades
Grading Scheme from a traditional learning environment[2]	High Distinction	85-100
	Distinction	75-84
	Credit	65-74
	Pass	50-64
	Fail	< 50
AdelaideX Course Grading Scheme	Pass	≥ 50
	Fail	< 50

Once the data is prepared according to the different grading schemes, linguistic features were extracted using the LIWC tool[3]. The feature extraction resulted in 93 linguistic features that exhibit different aspects of psychological and sociological perspectives. A comprehensive description of the psychological and sociological aspects of the LIWC features is given in Chapter 2.

Machine learning models were built for five different learner grades, as mentioned above. According to the results, the machine learning model with LIWC features resulted in

[2]https://www.adelaide.edu.au/policies/700
[3]https://liwc.wpengine.com/

predictive models that did not perform well due to its inability to distinguish the five different leaner grades. Thus, the grades were fine-tuned based on different categories to understand how well the model could perform in different grade settings.

After evaluating different grade settings, the feature space in the data was extended by incorporating 'time aspects' to understand whether or not continuous participation in MOOC discussion forums has an impact on learner grades. The aforementioned grade prediction model used the mean of linguistic features that were contributed over six weeks. The next set of models was built by incorporating the linguistic features for each week. For each linguistic feature, the overall average for each week was computed by taking the mean of the linguistic features for each student. This captured the learners' contributions in discussion forums over the weeks.

The models with both linguistic features (i.e., overall linguistic features) and time aspects (i.e., linguistic features per each course week) were re-trained to investigate the impact of the time aspects on the learners' grade prediction model and to understand whether they improved the accuracy of the model.

7.2.2 Rule Extraction

This section describes the process used to develop a set of rules to understand the relationship between the linguistic features and learner grades in detail. Apart from the decision tree classifier, the study implemented several other classification algorithms to examine the performance of the decision tree model compared to other classifiers. After building a series of machine learning models, a set of rules were derived from the decision tree classifier.

A decision tree is a flowchart-like tree structure consisting of features, decision rules and outcomes that are represented by internal nodes, branches and leaf nodes respectively. Like any other classifiers, the intention of a decision tree is to predict the target variable. The underlying concept behind this prediction is that the decision tree classifier learns the decision rules from the training data and predicts the target variable by applying these decision rules (Song & Ying, 2015).

A tree can be built by considering several options that can be used as nodes for splitting the data into subsets. To choose the best split, Attribute Selection Measures (Devi &

Nirmala, 2013) are considered when selecting the best attribute to split the data. The best attribute is used as a decision node and the data is further split into different subsets. This recursive partition is continued at each decision node.

An attribute selection measure is used when choosing the splitting condition that best splits the given data (Devi & Nirmala, 2013). It provides a heuristic for choosing the splitting criterion and determines how the data at a given node are split. 'Information Gain' was used as the attribute selection method in this thesis. Given a node, each feature can be considered as the splitting feature; however, the feature with the highest information gain is selected for the split.

Information gain is evolved from the notion of entropy, which is introduced by Claude Shannon in his information theory (Shannon, 1948). Entropy is defined as a measure of disorder or impurity. In a decision tree, entropy (E) is calculated at each decision node using the formula below:

$$E = -\sum_{i=1}^{c} Pi \log_2(Pi) \tag{7.1}$$

(Pi) - The proportion of the sample that belongs to class i for a particular node.

Entropy ranges between zero to one, it becomes zero when a node contains all the elements from a certain class: in other words, entire instances fall under either positive or negative classes. On the other hand, entropy gets its maximum value (i.e., one), when a node divides its instances equally within the target variable.

Information gain calculates the decrease in entropy. It calculates the reduction with its immediate lower level node in the given decision tree. In a decision tree, it measures the degree of information a feature gives about the target variable. During the split, a decision tree considers how to follow a particular split where maximum information gain is obtained. The information gain is represented below:

$$Gain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|Sv|}{|S|} \times Entropy(Sv) \tag{7.2}$$

- A - All possible values for attribute A.

- v - Any possible values of attribute A.
- S_v - The subset of S for whose attribute A has value v.
- $|S_v|$ - Number of elements in S_v
- |S| - Number of elements in S.

The entropy of the original collection S is represented by the first term (Entropy(S)) in the equation. The entropy value after S partitioned using attribute A is represented by the second term in the equation. More specifically, the second term is the sum of the entropies of each subset S_v, weighted by the fraction of examples (S_v/S) that belong to S_v (H. Wu, Zhang, Xie, Kuang, & Ouyang, 2013).

7.3 Results

This section describes the results obtained from grade prediction from linguistic-only features that are extracted from discussion forums, followed by a comprehensive analysis of rule extraction.

7.3.1 Grade Prediction using Discussion Forum Data

Several machine learning classifiers were implemented to predict learner grades with the linguistic features extracted from discussion forum posts. Classifiers were built with 10-fold cross validation with weighted average evaluation metrics using scikit-learn library (Pedregosa et al., 2011). Linguistic features that had been extracted using the LIWC tool were used to train and evaluate the classifier performance for different sets of learner grades, as mentioned in Table 7.1. Prior to model building, correlated features were removed and RFECV with 10-fold cross validation was performed to identify the optimal features.

7.3.2 Learner Grade Prediction with Linguistic-Only Features

The initial results show the grade predictions according to grading schemes from a traditional learning environment. According to the feature ranking with RFECV, 70 optimal

features are chosen as the best features to train the machine learning models to predict grades obtained in a traditional learning environment grade setting, as mentioned in Table 7.1. The classifiers were built using 10-fold cross validation. The results of the predictive model with linguistic features extracted from the LIWC tool are presented in Table 7.2.

TABLE 7.2: Classifier performance for a traditional learning environment grading scheme with five different learner grades

Classifier	Precision	Recall	F-Measure
Logistic regression	0.47	0.32	0.37
XG Boost	0.45	0.55	0.49
Decision Tree	0.49	0.28	0.33
Random Forest	0.45	0.54	0.48

The next study changed the grades to the AdelaideX course grading scheme used in the given course to understand whether it could perform better in predicting only two different levels of grades, namely Pass and Fail. The feature ranking with RFECV resulted in 66 optimal features as best features to train the machine learning models. Table 7.3 shows the performance of the machine learning model built with LIWC linguistic features to predict the learner grades according to the AdelaideX course grading scheme.

TABLE 7.3: Classifier performance for AdelaideX course grading scheme

Classifier	Precision	Recall	F-Measure	ROC Area
Logistic regression	0.66	0.64	0.65	0.68
XG Boost	0.66	0.66	0.66	0.73
Decision Tree	0.66	0.63	0.63	0.67
Random Forest	0.67	0.65	0.66	0.72

The overall results show the predictive models performed with less accuracy due to their inability to capture different score levels while predicting the fine-tuned grade settings that prevail in a traditional learning environment. A relatively better precision score has been obtained when predicting in accordance with the online course grading scheme. Therefore, in the next section, the study discusses how extending the feature space with 'time aspects' can improve the performance of the predictive models.

7.3.3 Learner Grade Prediction with Linguistic Features and Time Aspects

To improve the performance of the machine learning models, the feature space was extended by integrating 'time aspects'. In other words, weekly contributions of linguistic features by each student were computed. The statistics reveal that an average of 2040 posts were posted each week by learners. Among these posts an average of 1466 posts were posted by pass-grade learners whereas 573 posts were posted by fail-grade learners.

Classifiers were retrained and evaluated along with feature engineering using RFECV. The RFECV with 10-fold cross validation resulted in 239 features for predicting the grades according to the AdelaideX course grading scheme.

Table 7.4 presents the results obtained for classifiers built with linguistic features along with the time aspects.

TABLE 7.4: Classifier performance for AdelaideX course grading scheme with linguistic features along with time

Classifier	Precision	Recall	F-Measure	ROC Area
Logistic regression	0.73	0.69	0.69	0.75
XG Boost	0.70	0.69	0.69	0.76
Decision Tree	0.72	0.68	0.68	0.73
Random Forest	0.73	0.69	0.69	0.76
AdaBoost	0.69	0.69	0.69	0.75

Hyper parameter tuning was carried out as mentioned in Section 5.2.3. The Grid search hyper parameter tuning technique was performed to identify the best set of parameters where the classifier performs at its best. The parameter setting for the Random Forest classifier is given below:

criterion: 'entropy'

max_depth: 10

max_features: 'auto'

min_samples_leaf: 4

min_samples_split: 10

n_estimators: 400

7.3.4 Decision Rules

After implementing several machine learning classifiers, the decision tree machine learning model was selected to extract a set of rules that can have practical implications. The decision tree built from the decision tree classifier is presented in Appendix (see Appendix B).

The hyper-parameters of the decision tree model were tuned for early stopping of the growth of the decision tree. Actions such as pruning (e.g. limiting the depth of tree during the grid search) and evaluating the training and testing accuracy on each fold were conducted to prevent over-fitting in the model.

The entropy at each decision node is calculated using equation 7.1. The significant decision rules that are extracted from the decision tree are presented in Figures 7.2 to 7.26. These decision rules are presented in two formats as follows:

1. Overall mean value of a linguistic feature.
   <Linguistic Feature>

2. Mean value of a linguistic feature obtained on a particular week.
   Week<Number>_<Linguistic Feature>

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic >  0.50
|   |   |--- Week6_AllPunc >  5.72
|   |   |   |--- power <= 2.69
|   |   |   |   |--- weights: [0, 75] class: 1
```

FIGURE 7.2: Rule set 1 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic <= 0.50
|   |   |--- Week1_tentat <= 0.34
|   |   |   |--- Week6_tentat >  1.19
|   |   |   |   |--- Apostro <= 0.59
|   |   |   |   |   |--- weights: [0, 46] class: 1
```

FIGURE 7.3: Rule set 2 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic <= 0.50
|   |   |--- Week1_tentat >  0.34
|   |   |   |--- Week3_AllPunc >  1.25
|   |   |   |   |--- weights: [0, 89] class: 1
```

FIGURE 7.4: Rule set 3 extracted from Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic >  0.50
|   |   |--- Week5_discrep <= 3.53
|   |   |   |--- Week1_discrep <= 0.04
|   |   |   |   |--- Week2_Period >  3.88
|   |   |   |   |   |--- weights: [0, 101] class: 1
```

FIGURE 7.5: Rule set 4 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic >  0.50
|   |   |--- Week5_discrep <= 3.53
|   |   |   |--- Week1_discrep >  0.04
|   |   |   |   |--- weights: [0, 225] class: 1
```

FIGURE 7.6: Rule set 5 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|    |--- Week2_Authentic <= 0.50
|    |    |--- Week1_tentat >  0.34
|    |    |    |--- Week3_AllPunc <= 1.25
|    |    |    |    |--- Week5_Authentic >  17.02
|    |    |    |    |    |--- weights: [1, 45] class: 1
```

FIGURE 7.7: Rule set 6 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|    |--- Week4_Authentic >  0.50
|    |    |--- Week6_AllPunc <= 5.72
|    |    |    |--- anx >  0.16
|    |    |    |    |--- Week1_WC >  2.50
|    |    |    |    |    |--- Week4_adj >  3.27
|    |    |    |    |    |    |--- weights: [3, 73] class: 1
```

FIGURE 7.8: Rule set 7 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|    |--- Week4_Authentic <= 0.50
|    |    |--- Week3_compare <= 0.33
|    |    |    |--- Week2_i <= 2.30
|    |    |    |    |--- Week6_adverb >  0.62
|    |    |    |    |    |--- Dic <= 90.07
|    |    |    |    |    |    |--- Week1_social >  6.16
|    |    |    |    |    |    |    |--- weights: [3, 58] class: 1
```

FIGURE 7.9: Rule set 8 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|    |--- Week4_Authentic <= 0.50
|    |    |--- Week3_compare <= 0.33
|    |    |    |--- Week2_i <= 2.30
|    |    |    |    |--- Week6_adverb >  0.62
|    |    |    |    |    |--- Dic <= 90.07
|    |    |    |    |    |    |--- Week1_social <= 6.16
|    |    |    |    |    |    |    |--- you >  0.27
|    |    |    |    |    |    |    |    |--- weights: [3, 42] class: 1
```

FIGURE 7.10: Rule set 9 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|    |--- Week4_Authentic <= 0.50
|    |    |--- Week3_compare <= 0.33
|    |    |    |--- Week2_i <= 2.30
|    |    |    |    |--- Week6_adverb <= 0.62
|    |    |    |    |    |--- Week6_Analytic <= 0.50
|    |    |    |    |    |    |--- shehe <= 0.74
|    |    |    |    |    |    |    |--- Week1_discrep >  0.62
|    |    |    |    |    |    |    |    |--- verb >  19.20
|    |    |    |    |    |    |    |    |    |--- Week1_negemo >  0.42
|    |    |    |    |    |    |    |    |    |    |--- weights: [57, 5] class: 0
```

FIGURE 7.11: Rule set 10 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic <= 0.50
|   |   |--- Week3_compare >  0.33
|   |   |   |--- Week6_reward >  0.21
|   |   |   |   |--- weights: [5, 55] class: 1
```

FIGURE 7.12: Rule set 11 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic <= 0.50
|   |   |--- Week1_tentat <= 0.34
|   |   |   |--- Week6_tentat <= 1.19
|   |   |   |   |--- Week5_Comma <= 6.59
|   |   |   |   |   |--- focuspast <= 2.98
|   |   |   |   |   |   |--- adj >  4.88
|   |   |   |   |   |   |   |--- weights: [5, 46] class: 1
```

FIGURE 7.13: Rule set 12 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic >  0.50
|   |   |--- Week5_discrep <= 3.53
|   |   |   |--- Week1_discrep <= 0.04
|   |   |   |   |--- Week2_Period <= 3.88
|   |   |   |   |   |--- weights: [6, 51] class: 1
```

FIGURE 7.14: Rule set 13 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic <= 0.50
|   |   |--- Week3_compare <= 0.33
|   |   |   |--- Week2_i <= 2.30
|   |   |   |   |--- Week6_adverb >  0.62
|   |   |   |   |   |--- Dic <= 90.07
|   |   |   |   |   |   |--- Week1_social <= 6.16
|   |   |   |   |   |   |   |--- you <= 0.27
|   |   |   |   |   |   |   |   |--- auxverb >  9.03
|   |   |   |   |   |   |   |   |   |--- weights: [6, 44] class: 1
```

FIGURE 7.15: Rule set 14 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic <= 0.50
|   |   |--- Week3_compare <= 0.33
|   |   |   |--- Week2_i >  2.30
|   |   |   |   |--- Week2_WPS >  9.71
|   |   |   |   |   |--- compare >  2.38
|   |   |   |   |   |   |--- adj <= 5.32
|   |   |   |   |   |   |   |--- weights: [97, 13] class: 0
```

FIGURE 7.16: Rule set 15 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic >  0.50
|   |   |--- Week6_AllPunc >  5.72
|   |   |   |--- power >  2.69
|   |   |   |   |--- weights: [5, 38] class: 1
```

FIGURE 7.17: Rule set 16 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic <= 0.50
|   |   |--- Week1_tentat <= 0.34
|   |   |   |--- Week6_tentat <= 1.19
|   |   |   |   |--- Week5_Comma >  6.59
|   |   |   |   |   |--- weights: [6, 41] class: 1
```

FIGURE 7.18: Rule set 17 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic <= 0.50
|   |   |--- Week3_compare >  0.33
|   |   |   |--- Week6_reward <= 0.21
|   |   |   |   |--- Week2_WC >  25.00
|   |   |   |   |   |--- Week2_Authentic <= 31.18
|   |   |   |   |   |   |--- cause <= 2.06
|   |   |   |   |   |   |   |--- weights: [6, 43] class: 1
```

FIGURE 7.19: Rule set 18 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic <= 0.50
|   |   |--- Week1_tentat <= 0.34
|   |   |   |--- Week6_tentat <= 1.19
|   |   |   |   |--- Week5_Comma <= 6.59
|   |   |   |   |   |--- focuspast <= 2.98
|   |   |   |   |   |   |--- adj <= 4.88
|   |   |   |   |   |   |   |--- cause >  2.97
|   |   |   |   |   |   |   |   |--- weights: [6, 37] class: 1
```

FIGURE 7.20: Rule set 19 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic <= 0.50
|   |   |--- Week1_tentat <= 0.34
|   |   |   |--- Week6_tentat >  1.19
|   |   |   |   |--- Apostro >  0.59
|   |   |   |   |   |--- weights: [6, 37] class: 1
```

FIGURE 7.21: Rule set 20 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic <= 0.50
|   |   |--- Week3_compare <= 0.33
|   |   |   |--- Week2_i <= 2.30
|   |   |   |   |--- Week6_adverb >  0.62
|   |   |   |   |   |--- Dic <= 90.07
|   |   |   |   |   |   |--- Week1_social <= 6.16
|   |   |   |   |   |   |   |--- you <= 0.27
|   |   |   |   |   |   |   |   |--- auxverb <= 9.03
|   |   |   |   |   |   |   |   |   |--- article <= 9.62
|   |   |   |   |   |   |   |   |   |   |--- weights: [9, 51] class: 1
```

FIGURE 7.22: Rule set 21 extracted from the Decision Tree

```
|--- Week5_Analytic >  0.50
|   |--- Week2_Authentic >  0.50
|   |   |--- Week5_discrep >  3.53
|   |   |   |--- weights: [11, 64] class: 1
```

FIGURE 7.23: Rule set 22 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic >  0.50
|   |   |--- Week6_AllPunc <= 5.72
|   |   |   |--- anx >  0.16
|   |   |   |   |--- Week1_WC >  2.50
|   |   |   |   |   |--- Week4_adj <= 3.27
|   |   |   |   |   |   |--- weights: [10, 45] class: 1
```

FIGURE 7.24: Rule set 23 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic <= 0.50
|   |   |--- Week3_compare >  0.33
|   |   |   |--- Week6_reward <= 0.21
|   |   |   |   |--- Week2_WC <= 25.00
|   |   |   |   |   |--- WPS <= 28.36
|   |   |   |   |   |   |--- achieve >  2.81
|   |   |   |   |   |   |   |--- Week1_cogproc <= 7.34
|   |   |   |   |   |   |   |   |--- drives >  11.95
|   |   |   |   |   |   |   |   |   |---  weights: [59, 14] class: 0
```

FIGURE 7.25: Rule set 24 extracted from the Decision Tree

```
|--- Week5_Analytic <= 0.50
|   |--- Week4_Authentic <= 0.50
|   |   |--- Week3_compare <= 0.33
|   |   |   |--- Week2_i >  2.30
|   |   |   |   |--- Week2_WPS >  9.71
|   |   |   |   |   |--- compare <= 2.38
|   |   |   |   |   |   |--- Week2_focuspast <= 5.71
|   |   |   |   |   |   |   |--- space >  4.74
|   |   |   |   |   |   |   |   |--- Week2_reward <= 1.45
|   |   |   |   |   |   |   |   |   |--- weights: [76, 19] class: 0
```

FIGURE 7.26: Rule set 25 extracted from the Decision Tree

Those linguistic features that are extracted within the aforementioned 25 rules are explained below.

TABLE 7.5: Explanation for linguistic features

Feature	Description
Analytic	Measures the thinking style of an individual. It captures whether an individual expresses low analytical thinking (personal, or narrative) or higher analytical thinking (logical and hierarchical) through use of words.
Adverb	The total number of adverbs that are presented in the forum posts.
Adjectives (adj)	The total number of adjectives that are presented in the forum posts.
Punctuation (AllPunc)	Total punctuation used in the given text such as periods/full-stops, question marks and quotation marks.
Apostrophes (Apostro)	The total number of apostrophes used in the discussion forum post.
Authentic	Authenticity measures to what extent the language used in discourse is personal and self-revealing, rather than detached and guarded.
Anxiety (Anx)	Measures the level of anxiety expressed in the discourse through words such as *worried*, or *fearful*.
Article	The total number of articles used in the discussion forum post.
Achieve	Calculates the words that express the references to success and failure.
Auxiliary verb (Auxverb)	Calculates the auxiliary verbs used in the discourse.

Compare	Reflects the comparatives used by the student in the discussion forum posts to compare one entity with another with words such as *like*, *bigger* and *smaller*.
Comma	The total number of commas used in the discussion forum post.
Cognitive process (Cogproc)	Reflects the cognitive process expressed in the discourse.
Cause	Calculates the causation words used in the forum post, such as *because*.
Dictionary (Dic)	The total number of dictionary words used in the discourse.
Drives	Captures the needs and motives in the discourse. The overarching dimension is a combination of Achievement, Affiliation, Power, Reward and Risk.
Discrepancy (Discrep)	Measures the discrepancies expressed in the discourse.
Focuspast	Reflects the time orientation variable that references past events.
First person singular pronoun (i)	Represents the first person used in discussion forum posts.
Negative emotions (negemo)	Measures the negative emotions expressed in the discourse.
Period	Represents the total number of periods used in the discussion forum post.
Power	References the status and social hierarchies used in the forum post.
Reward	Measures the positive goals and approaches in the discourse.

Second person pronoun (you)	The total number of second person pronouns used in discussion forum posts.
Social	Measures the social relationships expressed in the discourse.
Space	Captures the relativity measures in the discourse.
Tentative	Tentativeness of the text is captured through words such as *maybe*, *perhaps*. The tentative language cues show whether an individual is uncertain or insecure about the topic they discuss. Moreover, it can also suggest that an individual has not yet processed the information.
Third person pronoun (shehe)	Represents the third person voice used in discussion forum posts.
Verb	The total number of verbs that are presented in the forum posts.
Word count (WC)	The total word count (WC) used in the given discussion forum post.
Words per Sentence (WPS)	Calculates the number of Words per Sentence used in the discussion forum posts.

In order to have practical implications, and to present a simple set of rules to the learning community, the rules were ranked according to the entropy and information gain. The rules that have a lower entropy are given priority. In other words, the rules with lower entropy are considered as the top-most priority. The individual decision rules along with their prediction probabilities are presented in Table 7.6. The rule sets are based on the descending order of entropy. Each decision rule in Table 7.6 is elaborated from Figures 7.2 to 7.26. The definitions of the linguistic features in these rules presented in Table 7.5.

Table 7.6: Decision Rules

Decision Rules	Majority	Probability
Rule 1	Pass	100
Rule 2	Pass	100
Rule 3	Pass	100
Rule 4	Pass	100
Rule 5	Pass	100
Rule 6	Pass	97
Rule 7	Pass	96
Rule 8	Pass	95
Rule 9	Pass	93
Rule 10	Fail	92
Rule 11	Pass	92
Rule 12	Pass	90
Rule 13	Pass	89
Rule 14	Pass	88
Rule 15	Fail	88
Rule 16	Pass	88
Rule 17	Pass	87
Rule 18	Pass	87
Rule 19	Pass	86

Rule 20	Pass	86
Rule 21	Pass	85
Rule 22	Pass	85
Rule 23	Pass	82
Rule 24	Fail	81
Rule 25	Fail	80

The study extracted the top 25 rule sets according to the entropy calculations, whereby these rules can predict the majority of the grade with more than 80% probability. The rule sets show how the contributions of these linguistic features in discussion forums affect learner grades. For example, if all the 'if statements' are satisfied in the code below, it ensures the student is more likely to Pass (Rule 5).

```
if Week5_Analytic > 0.5:
     if Week2_Authentic > 0.5:
        if Week5_discrep <= 3.53:
            if Week1_discrep > 0.04:
                        return 'Pass'
```

Similarly, if all the 'if-statements' are satisfied in the code below, it ensures the student is more likely to Fail with 92% probability (Rule 10).

```
if Week5_Analytic <= 0.5:
     if Week4_Authentic <= 0.5:
        if Week3_compare <= 0.33:
            if  Week2_i <= 2.3:
               if Week6_adverb <= 0.62:
                   if Week6_Analytic <= 0.5:
                       if shehe <= 0.74:
                           if Week1_discrep >  0.62:
                               if verb >  19.20:
                                   if Week1_negemo >  0.42:
                                           return 'Fail'
```

Analysing the rule sets shows that 'Analytic' contributions made at Week 5 have the highest information gain, demonstrating the linguistic feature's (i.e., analytical thinking) ability to predict the learner grades efficiently, based on the value it possesses during the discourse. According to the definition by Pennebaker (2015), 'analytic' measures the thinking style of an individual. The results show that if the overall level of analytical thinking expressed in week 5 forum posts (i.e., towards the end of course) is greater than 0.5, students are more likely to obtain a Pass grade in the course. It demonstrates that a student who expresses analytical ability during the end of the course has a high chance of obtaining a Pass grade. However, the remaining contribution of each feature also has an impact on predicting the probability of a Pass grade. In the rule set below, if all the statements are satisfied, the probability of the students obtaining a Pass grade is 100%.

```
if Week5_Analytic > 0.5:
    if Week2_Authentic <= 0.5:
        if Week1_tentat > 0.34:
            if Week3_AllPunc > 1.25:
                    return  'Pass'
```

Conversely, if all the statements below are satisfied, the probability of students obtaining a Pass grade is 85%.

```
if Week5_Analytic > 0.5:
    if Week2_Authentic > 0.5:
        if Week5_discrep > 3.53:
            return 'Pass'
```

The second highest information gain feature is 'Authentic', expressed at Week 2 in the sub-portion of the decision tree where the parent node Week 5 'Analytic' is greater than 0.5. Similarly, tentativeness and discrepancy are the next highest information gain features that have the ability to predict learner grades. According to the decision tree, if Week 2 'Authentic' is greater than 0.5, it always leads towards a set of significant rules with greater than 85% prediction probability in predicting the Pass grade. Since the 'Authentic' value expressed at Week 2 can be retrieved at an early stage of the course, the course instructors will have a high chance of understanding the learning stage of the students and be in a position to make timely interventions. Similarly, if the Week 2, 'Authentic' value is less than 0.5, all the significant rules can predict the learner grades as Pass.

The next interesting observation seen across the retrieved rule set is that any form of punctuation such as a period/full stop, apostrophes and overall punctuation can determine the probability of pass grade learners. For example, if Week 3 punctuation is greater than 1.25, passing the above three levels of conditions, the probability of obtaining a Pass grade is 100%, whereas if Week 3 punctuation is less than 1.25, the probability of obtaining a Pass grade ranges between 70% to 97%.

The next path of the tree shows the linguistic characteristics where the 'Analytic' expressed at Week 5 is less than 0.5. In this branch, the study observed both Pass and Fail as end results. This sub portion of the decision tree holds Week 4 'Authentic' as the 2^{nd} highest information gain feature. According to the results, if Week 4 'Authentic' is greater than 0.5, all the sub-branches lead to Pass grades with different probabilities. On the other hand, if 'Authentic' expressed at Week 4 is less than 0.5, the end results can be either Pass or Fail, Therefore it is important to look into the next highest information gain feature, which is 'Week 3 Compare'.

The results show the majority of the Fail grade classes can be seen on the path where Week 3 'Compare' is less than 0.33. However, there is a possibility of obtaining a Fail

grade even if Week 3 'Compare' is greater than 0.33, with 81% probability. This can be easily tracked by monitoring the Week 2 'Word counts' and overall 'Words per Sentence' expressed during the course.

As aforementioned, the majority of the Fail grades can be observed in the decision tree when Week 3 Compare obtains less than 0.33. In this decision tree path, personal pronouns ('I') expressed at Week 2 are the next highest information gain node that can determine pass and fail. According to the decision rules, if 'personal pronouns ('I')' expressed at Week 2 are greater than 2.3, the condition always leads to a Fail grade, with more than 80% probability. Similarly, if 'personal pronouns ('I')' in the Week 2 discourse are less than 2.3 and the next set of linguistic features demonstrates lesser values, except for 'negative emotions', then there is a higher chance that a student will obtain a Fail grade at the end of the course.

It is important to understand if a learner does not participate in the forum at a particular week, zero value is assigned for all the linguistic features in that week. Therefore, these linguistic behaviours can predict the learners' final course grade even without learners' participation in the discussion forum for some weeks. With these rule sets, instructors could track these predominant features expressed during different intervals to determine their interventions. For example, if an instructor identifies students who fail to use comparison language cues such as *different*, *dissimilar* and *easier* in their discourse, the instructor can intervene and guide learners to further progress as these learners have a high probability of failure. Similarly, often using personal pronouns ('I') at the early stage in discourse can be another language cue for instructors to identify learners who are likely to Fail in the final course assessment. Expressing negative emotions and using fewer words in the discourse during beginning of the course are some other early language cues that can predict Fail-grade learners. These language cues can also be an indication that instructors should intervene.

7.4 Discussion

Learner course grades can be considered as one indicator that prevails in any education setting to measure learning success. Course grades can describe learners' knowledge acquisition or how well a student has understood the course content.

On the other hand, course discussion forums provide a space where students can express thoughts, issues, knowledge, views and opinions; providing rich detail of their understanding and perspectives around content. The relationship between the language expressed in forums and learners' grades needs to be investigated to understand the impact of learners' language expression on course grades. Therefore, the aim of this study was to explore the correlations between linguistic features that are expressed in the discussion forums and learners' final course grades.

To begin with, predictive models were built under different sets of grading schemes (i.e., a Traditional learning environment and AdelaideX course grading scheme) to calculate the model score. Although the online learning platform uses Pass and Fail grades as a standard grading scheme, learner grades can be fine-tuned similar to traditional grade settings, as presented in Table 7.1. According to the results of learner grade prediction, it was found that the predictive model did not perform well in predicting fine-tuned learner grades, despite introducing new features to the feature space. Therefore, the study followed the AdelaideX course grading scheme used in the online learning platform for learner grade predictions. The results show learner grades can be predicted with a 69% F-measure. Although the model performance is sufficient to make predictions, the study can conclude several further factors related to the impact of linguistic features on performance prediction.

One of the highly possible reasons for the model performance is that the students' final course grades can also be influenced by several other factors. To begin with, several studies (Aljohani et al., 2019; Okubo et al., 2017) have investigated contextual features such as video watching time and click stream data. These contextual features can have an impact on grade predictions. However, the intention of this study is to analyse the impact of language expressions on course grades alone.

The next factor that could affect performance prediction is the level of participation in the discussion forums. To date, discussion forum participation is not mandatory in many MOOCs. Even though some students choose to participate, and their contributions are visible in discussion forums, optional participation can affect the performance of the model. In other words, Pass grade students might not fully contribute to the discussion forums due to their confidence level with the course content; similarly fail grade students might also not contribute as they are unfamiliar with the content of the course. These

participation levels can have an impact on predicting the performance of learners when using prediction models that are solely built upon the linguistic features extracted from discussion forums.

On the other hand, fluency in English language might be another factor that can impact the performance of a model prediction that uses linguistic features. Since the requirement of the course is to work in the English language, native English speakers can contribute a small amount of information in a descriptive way, whereas an informative post can be presented by non-native English speakers with a narrow range of vocabulary. These factors can affect the performance of a prediction model that uses linguistic-only features extracted from discussion forums.

Given the intention is to understand the relationship between language use and course grades, the study extracted rule sets from the decision tree classifier that are human friendly. Knowledge obtained from these rule sets can be converted into 'IF-THEN' rules which is the simplest form of representation. These rule sets help researchers and instructors to understand what kind of linguistic characteristics can be seen at different learner grades, even with optional participation. The top 25 rule sets were extracted in this study based on the entropy calculations.

According to the rule sets, 'Analytic' skill expressed in the discourse during the 5^{th} Week has the highest information gain in predicting the learner grades. Week 5 'Analytic' being greater than 0.5, it is the predominant factor that leads to a Pass grade. On the other hand, Fail grade learners express less 'Analytic' skill in Week 5 than the Pass grade learners. This shows the importance of analysing the Analytical skill towards the end of the course. However, linguistic behaviour expressed at the beginning of the course can also lead to predicting the eventual course grade of a learner. Early linguistic indications, such as 'negative emotions', 'personal pronouns ('I')' and 'word counts' expressed at beginning of the course and 'comparisons' expressed at the middle of the course, can be used for predicting learner grades.

Using the aforementioned 25 rule sets, instructors can guide the students to write discussion forum posts based on the Pass-grade learners' linguistic behaviours, such as asking learners to incorporate analytical skills into their discourse. This will stimulate the students to use their analytical skills in analysing the question posted by instructors/peers and eventually helps them to learn/understand the topics of the course. Simultaneously,

instructors can also formulate discussion forum questions that reflect certain linguistic features, such as asking learners to undertake comparisons rather than reflecting on one idea alone. Such guidance can help learners to navigate towards critical thinking or better knowledge acquisition in their learning process.

Since these rule sets are based on the decision tree, the accuracy of the decision tree model can be fine-tuned by expanding the feature space. To increase the accuracy of the decision tree model, the feature space can be expanded by incorporating features from Coh-metrix (McNamara et al., 2014), which captures additional linguistic features such as cohesion, connectivity and syntactic complexity; however, it lacks support for batch processing through the freely available version. Furthermore, meticulous feature engineering is important, as increasing the language features can result in a complex decision tree. Deep learning is an alternative approach to predict learner grades, but it is a black box approach that needs to be provided with large volumes of data. Provided that big data is available, and a black box approach is suitable, deep learning may have been a better choice to learn complex interactions among the features in finer detail to create a more accurate model for grade prediction. In addition, such a model can give us an opportunity to apply transfer learning for a new corpus. This shows that if the aim is only to predict learner grades without explaining why it happens (i.e., relationship between learners' linguistic behaviour in discussion forum with their course grades), deep learning models can be used.

The next section discusses how linguistic features extracted from learner discourse change with time and explores the differences in linguistics between two learner groups (i.e., Pass, and Fail).

7.5 Linguistic features with Time aspects

This study is designed to investigate linguistic expressions connected to different learner clusters; namely, high performers (i.e., Pass-grade learners) and low performers (i.e., Fail-grade learners) who are identified in section 7.2. The study investigates these linguistic expressions in connection with time to understand how they change throughout the course.

As mentioned in Chapter 5, several linguistic features were extracted using the Linguistic Inquiry and Word Count (LIWC) tool. These linguistic features were plotted against course duration to visualise how the different linguistic features that are exhibited through learner posts change with time between two learning groups (i.e., Pass-grade learners, and Fail-grade learners). The linguistic features were examined from the start to the end of the course to understand the changes happening during a course period. Linguistic features from two different components of the discussion forums (i.e., Comment Threads, Comments) were analysed individually to identify the level of contributions by learners for each element. Outliers that differed significantly from other observations were removed before the analysis. Mean (μ) values for each week were plotted in the graph for easy visualisation. μ_P and μ_F symbols represent the mean values of Pass-grade and Fail-grade learners respectively.

This study presents a set of features from each category given in the LIWC tool, as follows:

1. Summary Dimension
2. Function Words
3. Cognitive Process
4. Punctuation Marks
5. Time Orientation
6. Informal Language

7.5.1 Summary Dimension

The summary dimension in the LIWC tool presents summary aspects of the overall text, such as Words per Sentence (WPS), Analytical thinking and Clout. To begin with, two different graphs for total contributions and the mean distribution for WPS in the discussion forums by Pass and Fail groups are presented in Figure 7.27 and Figure 7.28 respectively. Figure 7.27 depicts how the linguistic measure 'Words per Sentence' of two different learner grades, namely Pass and Fail, change with time. According to Figure 7.27, it is evident that visualising learners' total contributions will always possess a

higher value for the Pass-grade learners than Fail-grade learners, as they hold a majority class in the entire learning group. Therefore, this study intends to analyse the average contributions of these two different learner groups in the discussion forum across time.

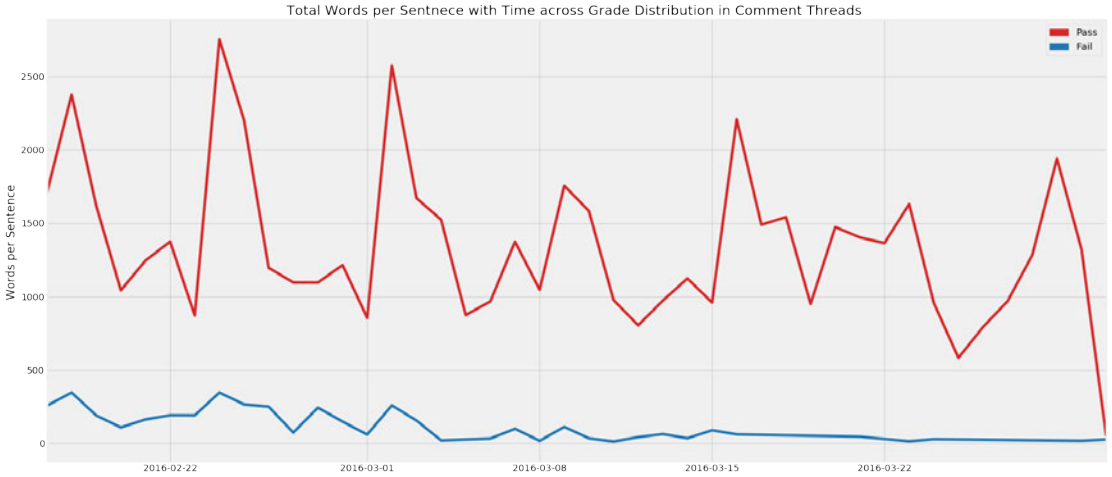


FIGURE 7.27: Total Words per Sentence in Comment Threads

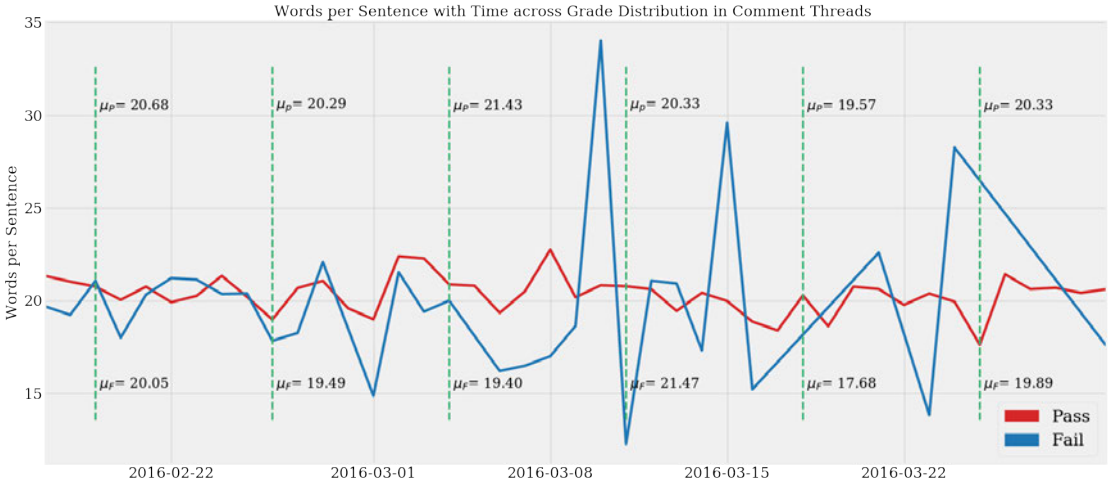


FIGURE 7.28: Average Words per Sentence in Comment Threads

Figure 7.28 shows how the linguistic measure 'Words per Sentence' of two different learner grades, namely Pass and Fail change over time in comment threads. Figure 7.28 shows that Pass-grade learners maintain approximately the same range of Words per Sentence in their writing throughout the course compared with the Fail-grade learners, whilst prominent fluctuations can be observed in Fail-grade learners during the end part of the course.

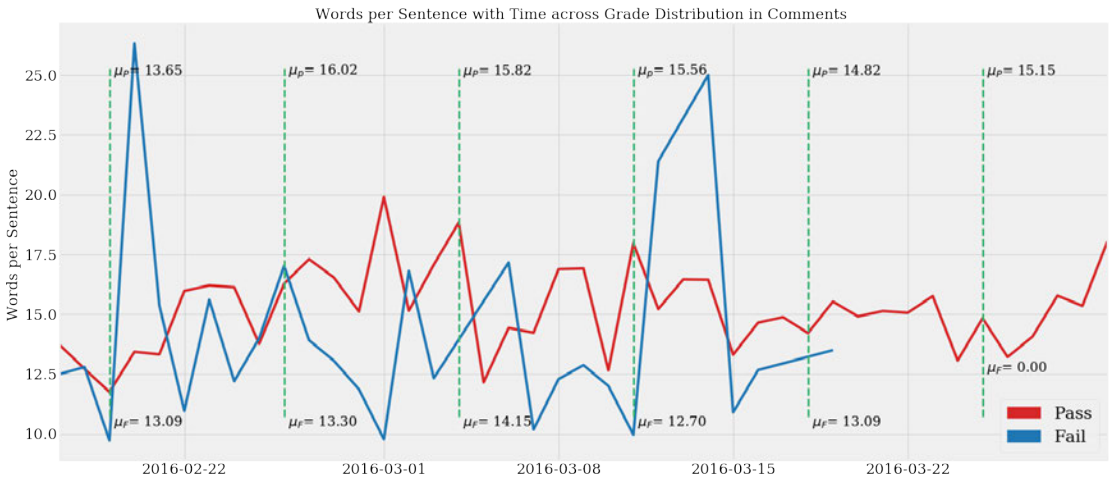


FIGURE 7.29: Average Words per Sentence in Comments

Figure 7.29 demonstrates the 'WPS' in the comments for these two different learning groups. Although Pass grade learners maintain a similar level of 'WPS', there is a rise at the beginning of the course. On the other hand, no contributions were observed in terms of 'Word Counts' and 'WPS', after removing the outliers in the comments section by the Fail grade learners.

The analysis shows that despite the fluctuations in 'WPS' that happens during the course, a Pass grade learner maintains a similar level of contribution throughout the course in both comment threads and comments compared with the Fail grade learners; whereas Fail-grade learners, apart from major fluctuations, were not contributing in peer interactions towards the end of the course.

Pennebaker et al. (2015) define the 'Analytical Thinking' in the LIWC tool as a measure of rational and hierarchical thinking capabilities. This is a measure to understand how a learner thinks and approaches different learning topics addressed during different time frames of the course.

Figure 7.30 and Figure 7.31 show the learners' analytical thinking across the course in comment threads and comments respectively. According to Figure 7.30, the analytical capabilities of Pass-grade learners maintain the same level throughout the course, with a drop towards the end in comment threads. On the other hand, Fail grade learners fluctuate from the middle of the course in the comment thread section. In the comments section, except for 5^{th} Week, Fail-grade learners possess a lower value than the Pass grade learners. Furthermore, their analytical thinking drops significantly at the last week of the course in the comment sections.

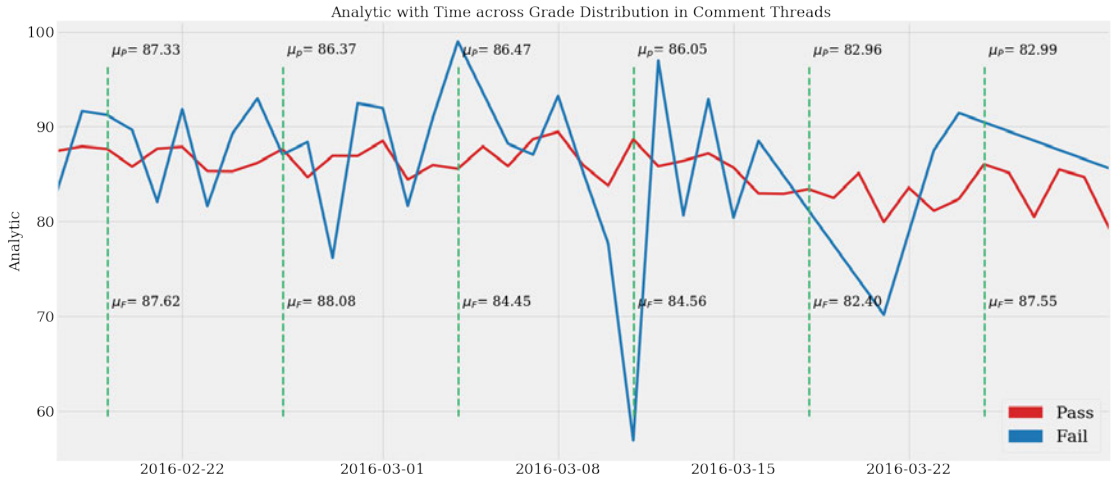


FIGURE 7.30: The Mean Usage of Analytical thinking in Comment Threads

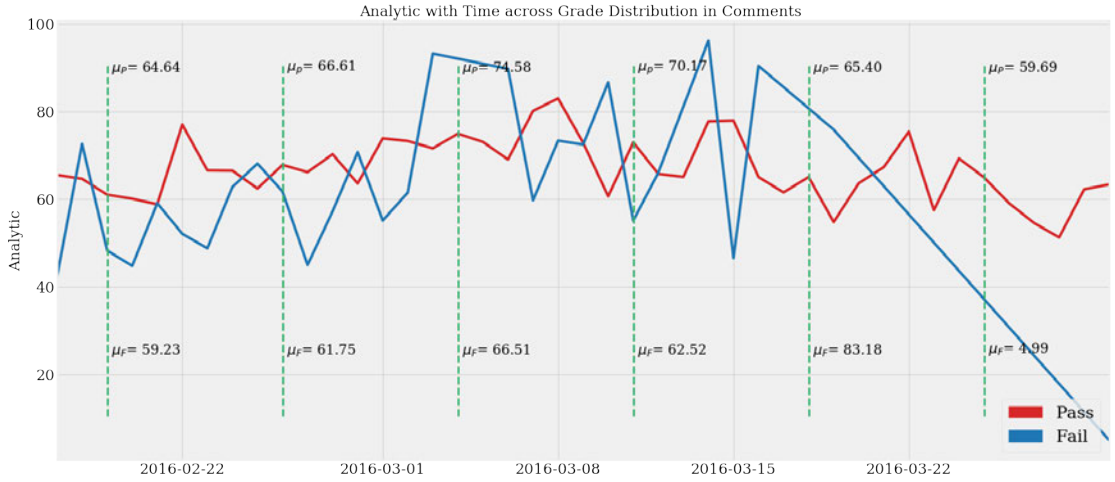


FIGURE 7.31: The Mean Usage of Analytical thinking in Comments

Since different learning topics have been discussed at each time interval, the high and low values for analytical thinking can be seen as an indication of the level of understanding of the course topics, as high values are aligned with logical and hierarchical thinking. Fail-grade learners mostly demonstrate a low analytical level in comments compared with Pass-grade learners, which can be explained by many possibilities such as their lack of understanding in interpreting the comments posted by their peers, or less time spent in peer interactions.

The Clout feature in the LIWC tool refers to the level of confidence in the discourse (J. W. Pennebaker, Booth, et al., 2015). High clout scores show an individual is more likely to maintain an authoritative persona, while low clout scores inclined to represent a more humble tone.

According to Figure 7.32 and Figure 7.33, Pass and Fail grade learners exhibit an increasing confidence level at the beginning of the course in comment threads. The clout score drops significantly at week 4 for Pass-grade learners and at the 5^{th} week for Fail-grade learners. Overall, Fail grade learners exhibit higher clout values than the Pass grade learners in their discourse. However, no peer interactions were observed for Fail-grade learners at the end of the course, except for the last day.

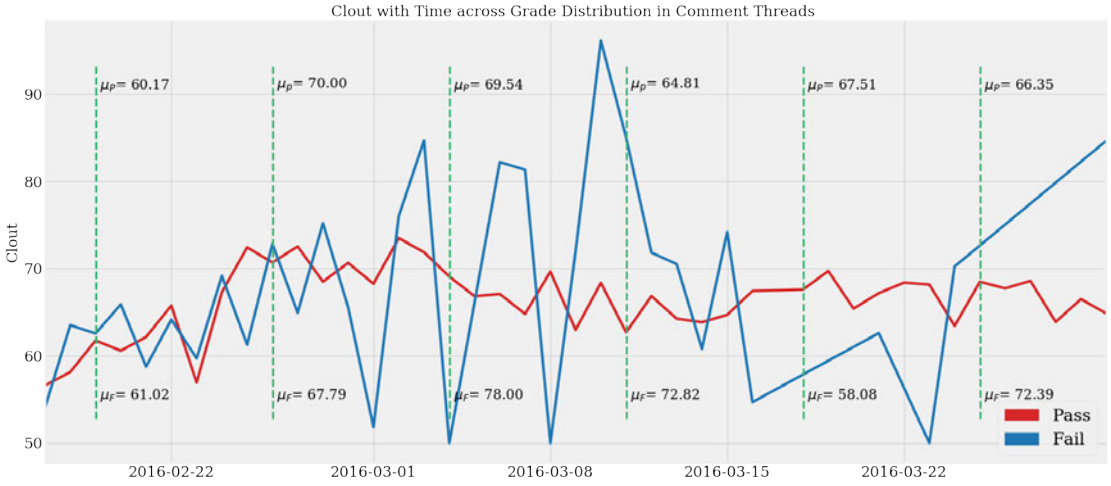


FIGURE 7.32: The Mean Usage of Clout in Comment Threads

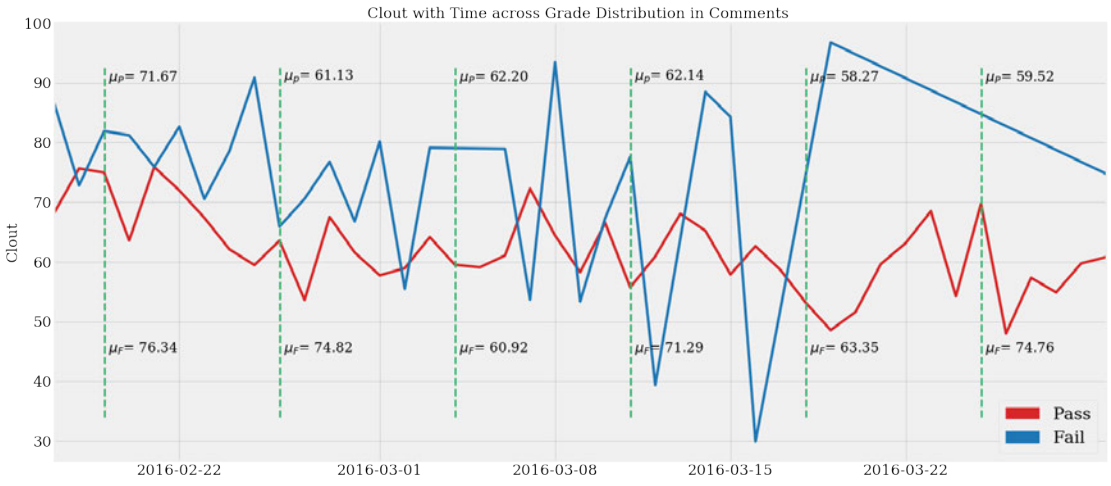


FIGURE 7.33: The Mean Usage of Clout in Comments

Analysing the overall data shows, few linguistic features demonstrate a sudden increase at the last week (i.e., only posted at 30^{th} March 2016) for Fail-grade learners as they tend to post more at the last day of the course. In contrast, Pass-grade learners are actively participating in peer interaction throughout the course.

7.5.2 Function Words

Function words in the LIWC tool contain measures such as pronouns and articles. Figure 7.34 and Figure 7.35 illustrate the use of personal pronouns ('I') in discussion forum posts with time in comment threads and comments respectively. According to Figure 7.34, use of personal pronouns drops during the middle of the course and increases at the end of the course for Pass-grade learners. On the other hand, Fail-grade learners also show major fluctuations, especially from the middle of the course.

According to Figure 7.35, pronouns usage by Pass grade learners are almost consistent until the middle of the course and increase towards the end; whereas fluctuations are observed throughout the course for Fail-grade learners in the comment section.

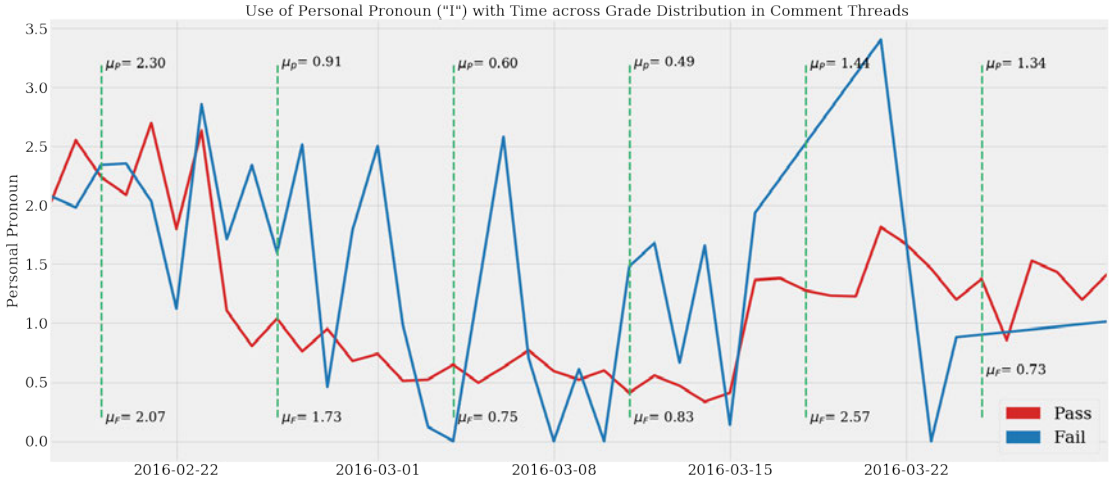


FIGURE 7.34: The Mean Usage of Personal Pronouns ('I') in Comment Threads

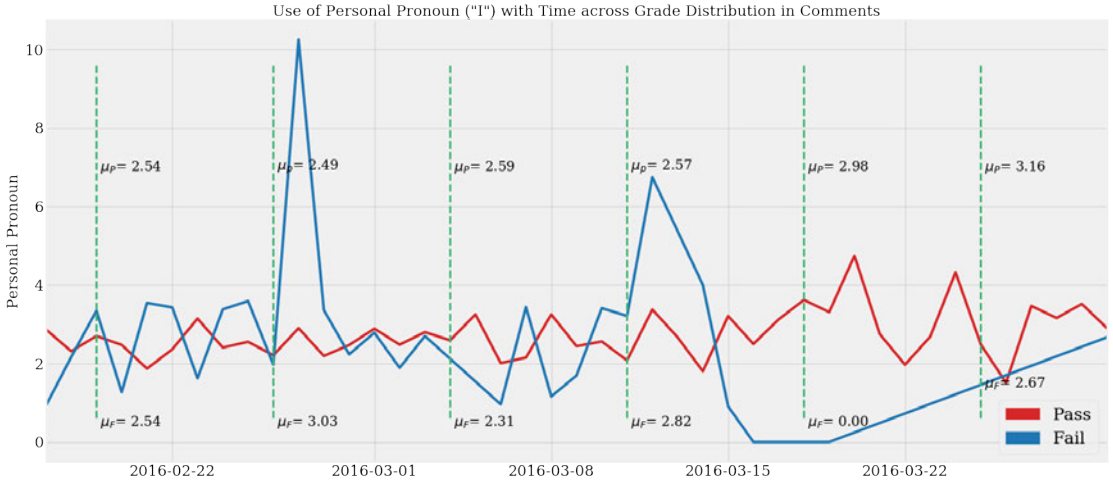


FIGURE 7.35: The Mean Usage of Personal Pronouns ('I') in Comments

7.5.3 Cognitive Process

The cognitive process in the LIWC tool is measured using various sub components, such as certainty, differentiation and causation. Figure 7.36 and Figure 7.37 show the changes in expressing 'differentiation' with time. According to the findings, the amount of cognitive processes reported in discussion forums for Pass-grade learners are generally high compared with those of the Fail-grade learners in the overall discussion forum, especially in the comments section. Analysing the analytic values in the comment section shows that a high value for Fail-grade learners is observed on the last day of the course in the comment section, whereas zero values are observed for other days in the final week.

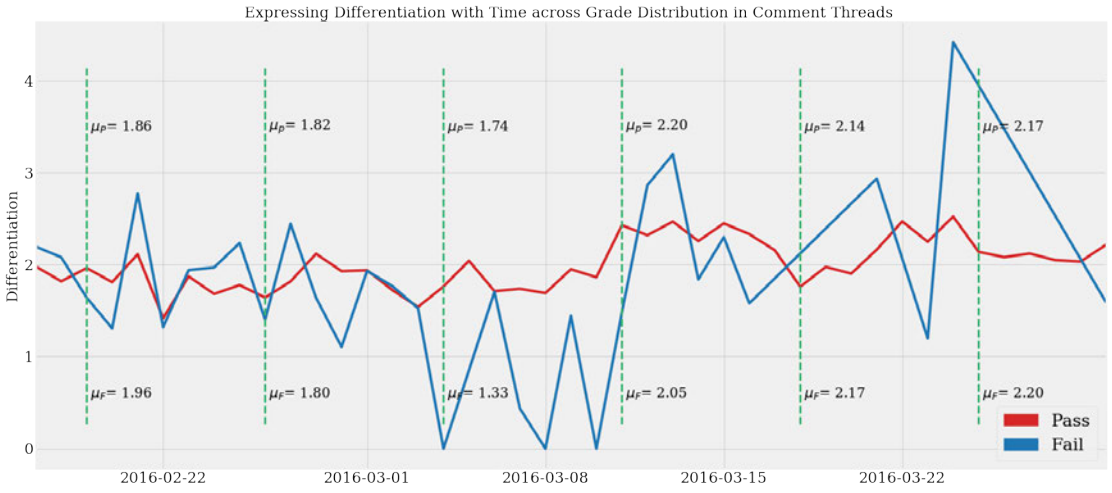


FIGURE 7.36: The Mean Usage of Differentiation in Comment Threads

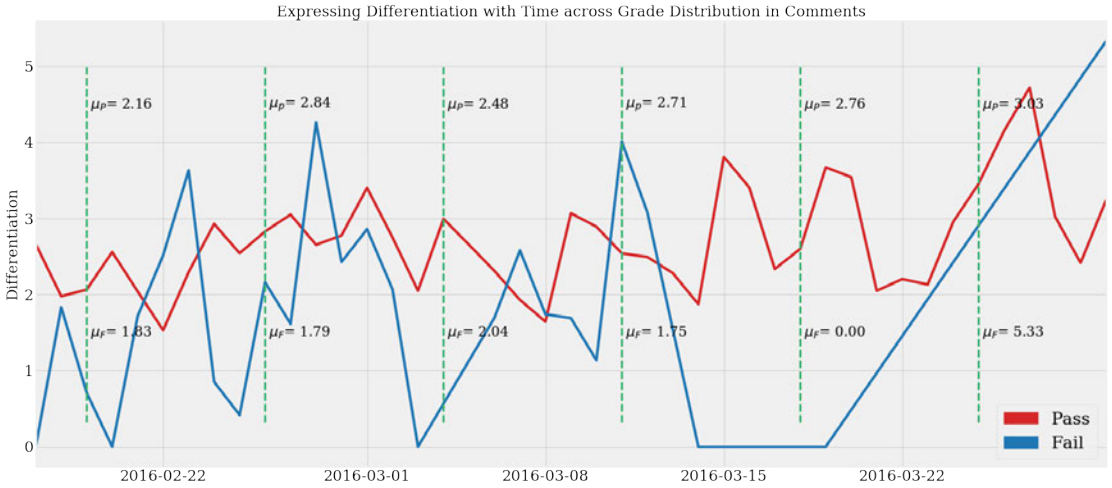


FIGURE 7.37: The Mean Usage of Differentiation in Comments

7.5.4 Punctuation Marks

Punctuation marks in the LIWC tool included items such as question mark, comma and period/full stop. According to the statistical analysis, the only significant punctuation in the discourse was the question mark.

The analysis shows an interesting pattern in the use of question marks across these two different learner grades. Figure 7.38 demonstrates how question marks have been used by these two different learners in comments. The results show that there is no significant use of question marks observed in the comment threads for both groups of learners. Similarly, no significant use of question marks was observed in comments for Pass-grade learners.

This overall analysis of question marks reveals the seeking behaviour of a learner. For Pass-grade learners, information-seeking in comment threads and comments is insignificant, whereas Fail-grade learners tend to seek information during peer interactions. This shows that though they have answered the question posted by the instructor in the comment threads, they have doubts about the information that has been posted by their peers.

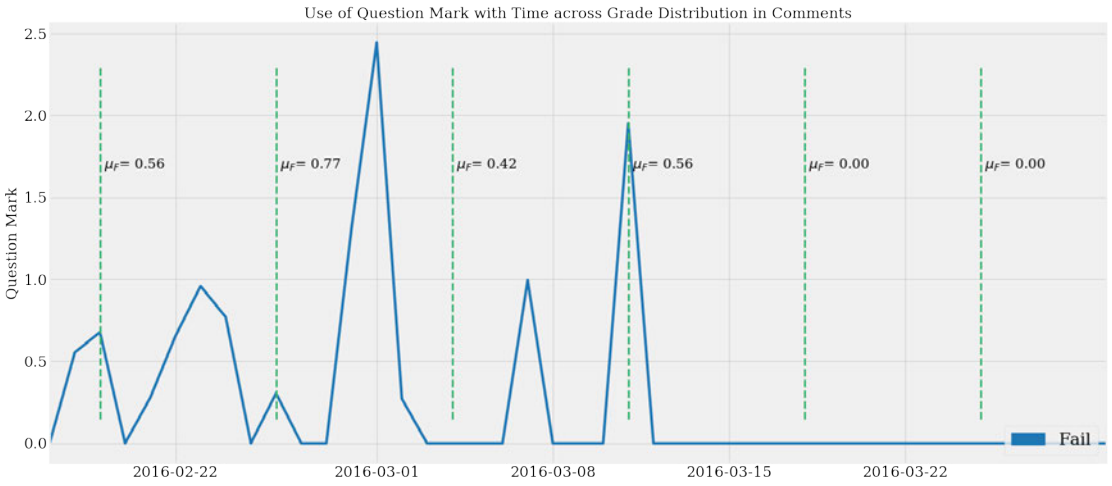


FIGURE 7.38: The Mean Usage of Question Marks in Comments

7.5.5 Time Orientation

The time orientation category in the LIWC tool consists of time orientation measures; namely, past, present and future orientation. Figure 7.39, Figure 7.40 and Figure 7.41

present how the references to these time orientations in discussion forum posts change with time in comment threads, while Figure 7.42, Figure 7.43 and Figure 7.44 depicts the time focus in comments.

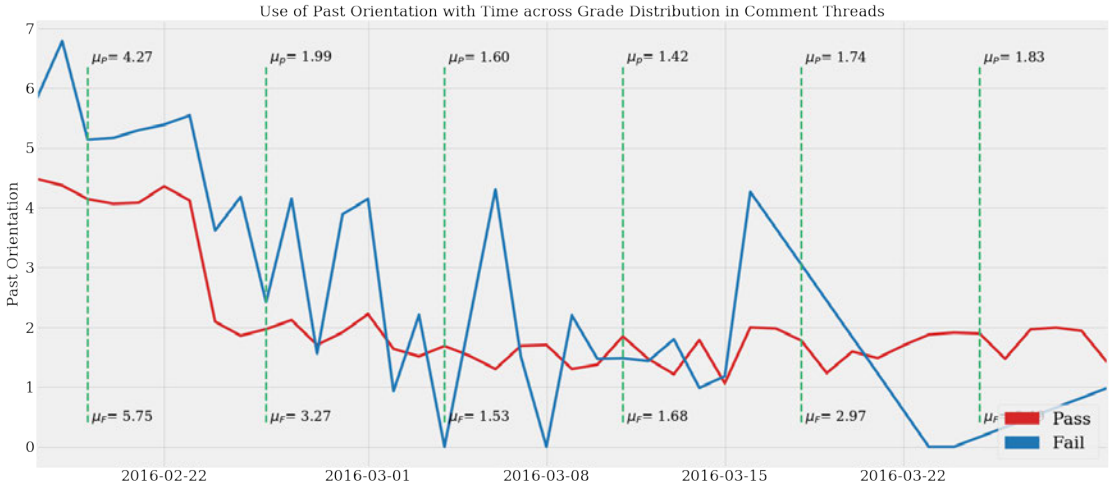


FIGURE 7.39: The Mean Usage of Past Orientation in Comment Threads

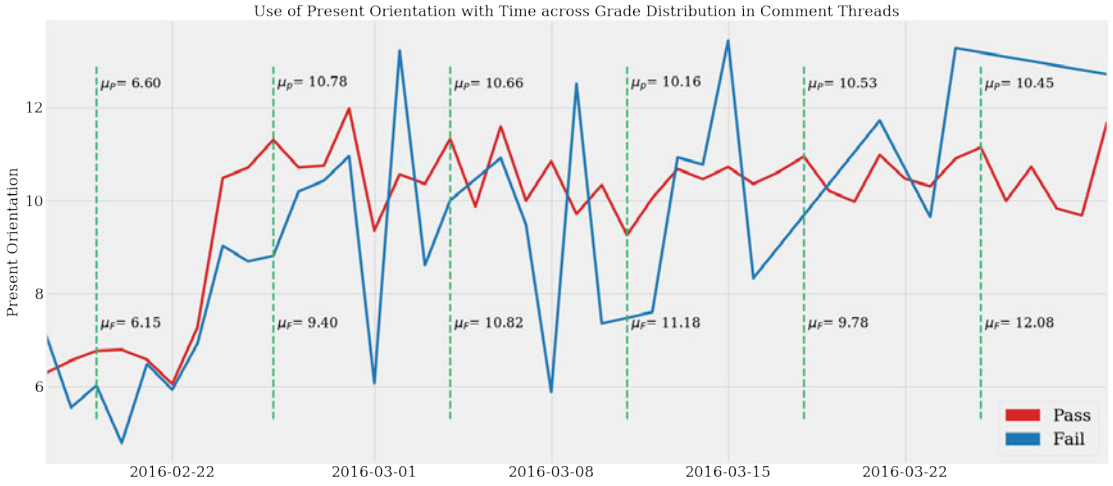


FIGURE 7.40: The Mean Usage of Present Orientation in Comment Threads

According to Figure 7.40 and Figure 7.43, references to present events in a discussion forum posted by both Pass-grade and Fail-grade learners are high compared with past and future references in their discourse. The results also show that Fail-grade learners tend to reduce past references in their comment thread posts and continue to increase their present and future references towards the end of the course. Conversely, the Pass-grade learners maintain almost the same level of references throughout the course in comment threads, despite a sudden increase in using present orientation and a decrease in past orientation at the beginning of the course, while a sudden increase in using future orientation at the middle of the course.

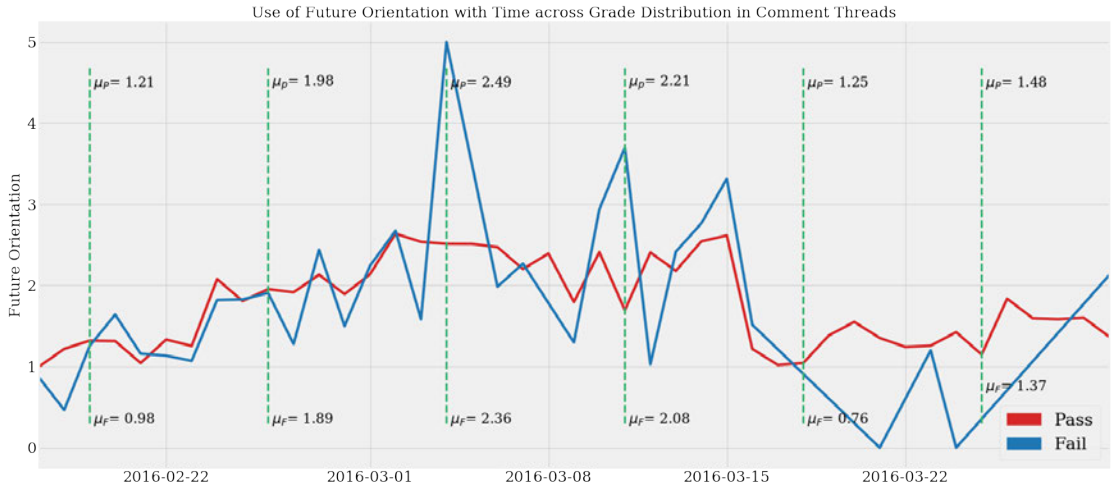


FIGURE 7.41: The Mean Usage of Future Orientation in Comment Threads

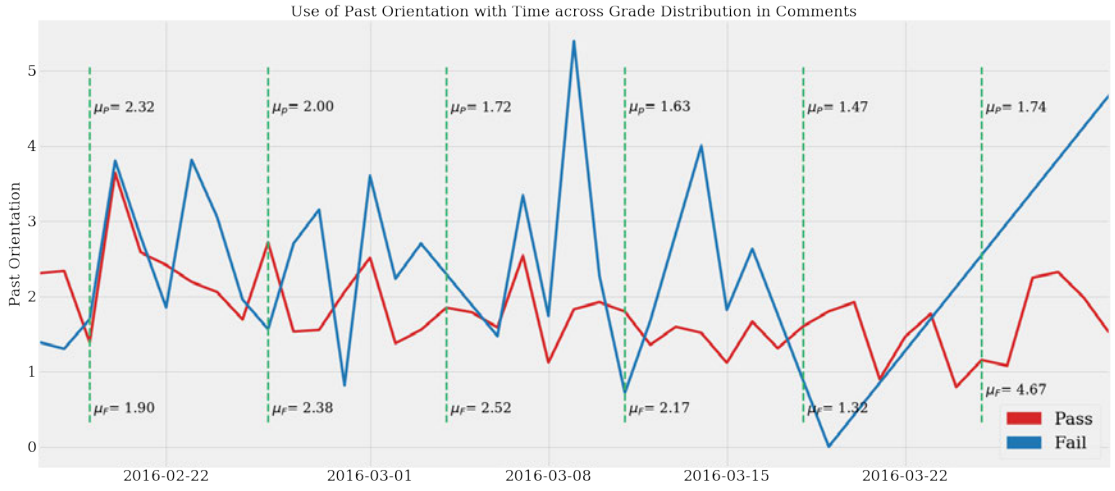


FIGURE 7.42: The Mean Usage of Past Orientation in Comments

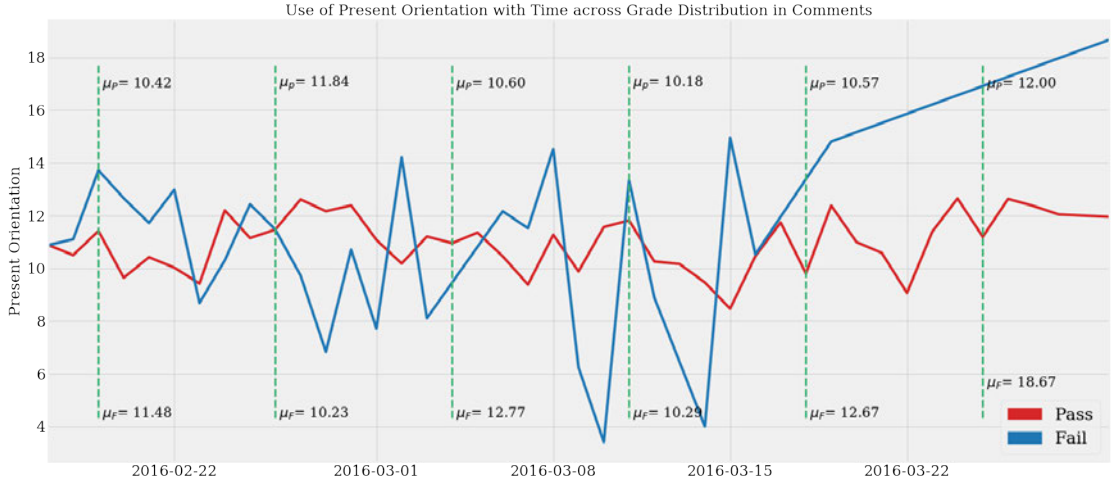


FIGURE 7.43: The Mean Usage of Present Orientation in Comments

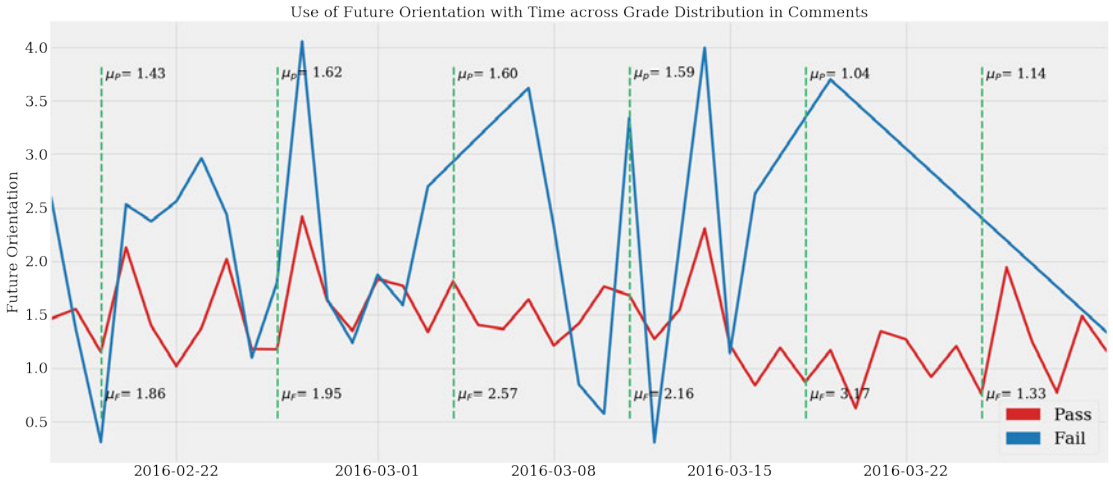


FIGURE 7.44: The Mean Usage of Future Orientation in Comments

7.5.6 Informal Language

Informal language in the LIWC tool refers to use of informal markers in the text, such as nonfluencies, filler words and net-speak words. Figure 7.45 and Figure 7.46 illustrate the average use of informal language in comment threads and comments respectively.

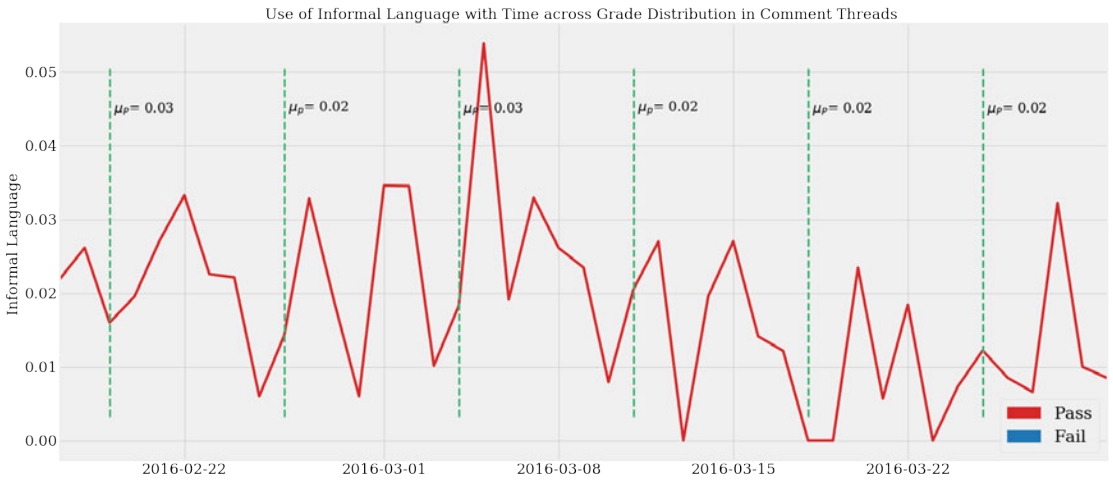


FIGURE 7.45: The Mean Usage of Informal Language in Comment Threads

According to Figure 7.45, it is evident that Pass-grade learners use informal language in comment threads; however, the values depict they have used minimal informal words, whereas Fail-grade learners possess a zero average throughout the course in comment threads. Conversely, Fail-grade learners use comparatively more informal language than Pass-grade learners in comments.

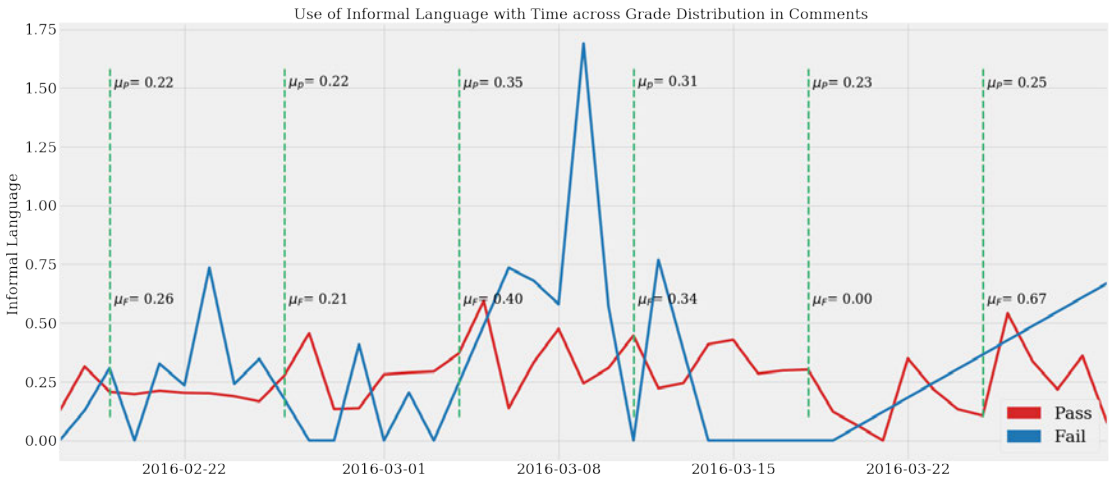


FIGURE 7.46: The Mean Usage of Informal Language in Comments

7.6 Linguistic Profiles

The objective of this study is to propose 'Linguistic Profiles' that help to identify and differentiate two different learner clusters: Pass-grade learners and Fail-grade learners. These linguistic profiles can act as a guide for understanding two different learner groups in a MOOC environment. This study has derived a set of linguistic features that are significantly different between these two learner groups. Linguistic features are derived from the LIWC tool that were used to build the grade prediction model. These linguistic features were computed at course-level. This dataset contained 4068 pass-grade learners and 2677 fail-grade learners. Moreover, the statistics reveal that an average of 2040 posts were posted each week.

To begin with a statistical analysis, the study performed a normality test to check whether the linguistic features follow a normal distribution. Since it is one of the underlying assumptions of statistical analysis, it is vital to perform a normality test. The study performed a normality test for each linguistic feature using the Shapiro–Wilk test.

According to the results, the study cannot use MANOVA (Multivariate Analysis of Variance) on the dataset as it violates the normality of the distribution. Therefore, the study used the Mann-Whitney U test (Mann & Whitney, 1947) to determine whether there is a statistically significant difference between the linguistic features of different learner grades. The assumptions of the Mann-Whitney U test were tested to ensure they are not violated. The results of the Mann-Whitney U test are presented in Table 7.7.

TABLE 7.7: Results of Mann-Whitney U test

Linguistic Feature	Grade	Mean Rank	Significance
Analytic	Pass	2362.32	0.000238
	Fail	2088.47	
Authentic	Pass	2453.65	0.003240
	Fail	2681.66	
Clout	Pass	2436.51	0.000081
	Fail	2742.03	
Dictionary	Pass	2370.70	0.000124
	Fail	2662.04	
Words per Sentence	Pass	2363.61	0.001276
	Fail	2121.57	
Personal Pronoun (i)	Pass	2319.55	0.000879
	Fail	2555.24	
Negation	Pass	2311.94	0.004528
	Fail	2118.98	
Personal Pronoun	Pass	2397.23	0.000014
	Fail	2728.61	
Differentiation	Pass	2417.42	0.020731
	Fail	2242.47	
Informal	Pass	2000.98	0.000653
	Fail	2117.99	

Focus Future	Pass Fail	2400.07 2555.34	0.038365
Focus Past	Pass Fail	2337.30 2744.41	$2.6563E^{-8}$
Focus Present	Pass Fail	2448.57 2285.68	0.034230

According to the statistical analysis, a set of features that significantly defines Pass-grade learners was identified in this study. The results show that analytical thinking for Pass-grade learners was statistically significantly higher than that for Fail-grade learners (p=0.00024). Analysing the discourse from Pass-grade learners shows that these learners usually write their discourse in a logical way, where the flow can be seen among sentences and they demonstrate hierarchical thinking in their discourse. The following are a few example posts posted by Pass-grade learners during the end of the course:

I have used Dropbox, Google Drive and Basecamp in the past as communication tools to collect and provide information to both internal and external stakeholders. I have recently started using Onedrive as the primary tool due to a request from management. All of the tools have been excellent and provide transparency to all stakeholders and allow all to see how the project is moving along and if required stages are being successfully completed. This is great for motivation and all can see how close we are to the project being completed successfully. The transparency also prevents unexpected surprises and stakeholders from not completing tasks and requirements as all parties roles and task are known to all. **- Pass-grade Learner**

In the three scenarios the project manager has to: Define the achievable goals to accomplish everything that needs to be done. Set up an initial and a final date to each goal in order to ensure the success of the project. Allocate resources to each goal, including the person responsible for each task, and the available budget to do so. **- Pass-grade Learner**

Similarly, the statistical analysis demonstrates that Pass-grade learners express differentiation significantly higher in their discourse than Fail-grade learners. Analysing the discourse shows, Pass-grade learners are able to differentiate entities in their discourse more effectively than Fail-grade learners using language cues such as 'but', 'differ' and 'rather'. The following is an example post from a Pass-grade learner who uses language cues to reflect differentiation:

I love Onedrive, I prefer it rather than Google Drive because I have the documents as they are, not just links. **- Pass-grade Learner**

The results also demonstrate that Pass-grade learners use significantly higher number of words per sentence and negation compared with the Fail-grade learners in their discourse. According to the literature, WPS are correlated with learning gain (Rosé et al., 2003; Core et al., 2003). Further it also demonstrate the linguistic complexity of the discourse. A study by Dowell et al. (2015) also identifies descriptive discourse as a discourse dimension among the students who perform better. Analysing their discourse shows that Pass-grade learners have used significant negation during their peer interactions through comments to share their opinions and knowledge on a topic, as follows:

What about finance? No money, no project. I believe that there is no phase within a project that is more important than another phase. Without one phase being developed all other phases fail, or don't work very well. **- Pass-grade Learner**

Conversely, Fail grade learners possess statistically significantly high values for clout, personal pronouns, informal language, authenticity and question marks in their discourse in comparison with Pass-grade learners.

Analysing the linguistic profile of Fail-grade learners shows that they have used personal pronouns ('I') more often than Pass-grade learners. The following is an example post from a Fail-grade learner who uses personal pronouns in their discourse:

The project that I have on top of my head is an induction for our new staff that i had to coordinate. The project was already initiated and planned when I joined the organization, and I had to take over the execution part of it as my colleague was sick. I found it challenging as I was not involved in the project from the start. I now am the

induction coordinator, and I am going to start working on the initiation part of the project from next week. The problem I am now facing is that I do not know where or how to start. **- Fail-grade Learner**

Similarly the results depict Fail-grade learners' Clout scores are significantly higher than those of the Pass-grade learners. Clout is defined as the level of confidence the author has on the given topic (J. W. Pennebaker, Booth, et al., 2015; Moore et al., 2021). It is a surprising finding that Clout scores are higher for Fail-grade learners. Analysing their discourse shows that Fail-grade learners convey their discourse more confidently, as follows:

Absolutely agree. First she needs to do some kind of "needs assessment" in order to have clear understanding of what she is expected to do. And as far as it is said that the clients have little idea on that, she can help them with hints driven from her previous experiences. Also she needs to consider the time frame and to offer services that are realistic to implement in that short time period. **- Fail-grade Learner**

On the other hand, the results show that Fail-grade learners often use dictionary words and informal language in their discourse. According to Tausczik and Pennebaker (2010), dictionary words are an indication of non-technical language. In other words, using fewer dictionary words reflects the number of technical terms used in their forum posts. Furthermore, studies have found that informal language, such as tentative words (e.g., maybe, perhaps and guess) and filler words (e.g., blah, I mean, and you know) in discourse is predominantly found in individuals who are uncertain about a topic (Tausczik & Pennebaker, 2010). Authenticity is another linguistic feature identified in the Fail-grade learner profile. High authenticity in the LIWC tool is associated with more honest, personal and disclosing text.

Finally, the results demonstrate that question marks have been used often by Fail-grade learners, demonstrating their help-seeking behaviour. The following example shows a series of questions asked by a Fail-grade learner.

Why the risk is considered in the Execution process and not at Planning process, when I answer "What risks do I need to consider?" in the execution process, is it not too late

to consider? I think it is already a cost (in worse scenario) to take up when the project is going in the not planned way or we considered it not an important risk at the beginning? **- Fail-grade Learner**

The aforementioned results were extracted from learners' discussion forum posts. Therefore, it is possible to foresee a course-grade by inspecting a learner's posts based on the aforementioned language cues. Furthermore, these results are only applicable when learners participate in discussions.

7.7 Discussion

The results of the 'Linguistic Profile' show the distinctive linguistic features that can define a learner profile. According to the results, a set of linguistic features have been identified to describe the learner profile of Pass-grade and Fail-grade learners in a MOOC learning environment.

Linguistic profiles demonstrate students who perform better (i.e., obtain Pass-grades) will have high analytical thinking than the lower performing students (i.e., obtain Fail-grades). According to Pennebaker et al. (2015), analytical thinking reflects the thinking style of a discourse. Higher values are associated with formal, logical and hierarchical thinking; while lower numbers demonstrate more informal, personal and narrative thinking.

The next significant linguistic feature found in Pass-grade learners' profile is differentiate; a category under the cognitive process of the LIWC tool. It was examined by mapping cognitive word use, which indicates the level of cognitive processing. Similarly, the statistical analysis confirms the Pass-grade learners hold statistically significantly high values for Words per Sentence, Negation and Focus present compared with Fail-grade learners.

The results confirm Fail-grade learners often use personal pronoun('I') in their discourse. A study (Atapattu et al., 2019) conducted to identify confusion in the education context show that the use of personal pronouns ('I') is higher among confused learners. Moreover, work by Rude et al. (2004) reveals individuals who are depressed use more first person

pronouns while expressing their thoughts. Therefore, personal pronouns ('I') can be an indicator of those learners who are likely to obtain Fail-grades at the end of the course.

Examining the clout score revealed Fail-grade learners express high levels of confidence (i.e., Clout) in their discourse compared with Pass-grade learners. The study (Moore et al., 2021) conducted to identify the relationship between clout and cognitive process reveals that differentiation was negatively associated with clout scores. The study emphasises the increased percentage of differentiation words is accompanied by decreased clout in posts. This finding aligns with the results of the linguistic profile, where differentiation is highly observed in Pass-grade learners' discourse, while clout is higher in Fail-grade learners' discourse.

The next set of features that define the learner profile of the Fail-grade learners is informal language, dictionary words and question marks. The study results also highlight Fail-grade learners often express their ideas using non-technical language. Furthermore, they also convey more informal language, confirming their uncertainty about course topics.

Temporal focus defines an individual's time perspective as the inclusiveness of their psychological past, present, and future (Lewin, 1942). Shipp et al. (2009) define the temporal focus as 'the extent to which people characteristically devote their attention to perceptions of the past, present, and future'. Shipp et al. (2009) also emphasise that a 'temporal focus profile is best represented by three main effects of past, current, and future focus'. Therefore, this study concluded that the temporal focus profile of Fail-grade learners is high reference to past and future events more than Pass-grade learners used in their discourse.

Fail-grade learners were mostly involved in explaining their past events related to project management, which is one of the questions posted in Week 1, and analysing their forum posts shows they have contributed less to analysing the case studies, given as the examples in the course. However, Pass-grade learners contributed equally to analysing the case studies, which is mostly reflected through present temporal focus rather than past focus.

Surprisingly, Fail-grade learners appear to be more authentic in their forum posts. Earlier research studies (Figueiredo, Soares, Vieira, Devezas, et al., 2020) have demonstrated that high authenticity does not always correspond to truth. Authentic discourse could

be more attractive and well accepted, without being true. Therefore, Fail-grade learners seem to adopt a more personal discourse.

7.8 Summary

This chapter presented a study on linguistics extracted from learners' discussion forum posts. The overarching aim of this study is to present a set of rules to the learning community to identify and distinguish learners who are likely to obtain Pass and Fail grades. This chapter described how a set of rules can be extracted using the decision tree algorithm. The first major contribution of this study is the rule set that explains the linguistic behaviours of Pass-grade and Fail-grade learners. The results from the study showed major rule sets can predict the learner grades with 80% probability, even with optional participation in some weeks. These rule sets can be applied to the subsequent offerings to ensure its validity. Moreover, this study also provides a detailed methodology that can be adopted by other courses from dissimilar domain to derive rule sets.

The study also visualised the linguistic features with time to understand how these two different learning groups have used different language cues in learner posts throughout the course. The study visualised these features on two different components of the discussion forum: Comment Thread and Comments. According to the analysis, the learners' contributions to particular linguistic features for these two different components differ significantly between the grades. For example, 'Questions Marks' were used by Fail-grade learners in the comments section, while no significant use of question marks was found for Pass-grade learners. Furthermore, analysing the discourse of these two different components shows that peer interaction of Fail-grade learners decreases with time, with zero contributions observed during the last week except for one day.

As the next contribution of this study, this chapter presented 'Linguistic Profiles' of two different learning groups (Pass and Fail) by describing the significant linguistic features that are reflected most of all in their discourse. The analysis shows, for example, that Pass-grade learners demonstrate more analytic and cognitive capabilities along with high usage of WPS, negations and present tense. Conversely Fail-grade learners exhibit help-seeking behaviour, along with high use of personal pronouns, clouts and dictionary words. The analysis also highlights the fact that a learner from any of these learning groups can

have a mixture of these features. It is important to closely observe the features that can mislead the instructors from identifying Fail-grade learners: for example, clout scores that reflect the confidence levels are high for Fail-grade learners.

With the findings of this study on linguistics, the next chapter presents the conclusion of this thesis by combining the findings observed in Chapter 5 and Chapter 6.

Chapter 8

Conclusion

The overall aim of this thesis was to investigate student learning in MOOCs by exploring the learner roles and linguistic expressions in discussion forums. Investigating the interactions in discussion forums can provide useful insights about students' learning as the discussion forum is a place where learners reflect and share their knowledge.

Different discussion forum structures require different types of analysis to extract informative data about students' learning. Analysing the discussion forum from the edX learning platform revealed several facts that might be useful for extracting information. This thesis uses discussion forum data from the 'Introduction to Project Management' course from the AdelaideX platform where forum participation is not mandatory. Furthermore, limited levels of discussion are observed in this discussion forum. In other words, different levels of discussions are stored only in two different types of objects in the discussion forum post data file (i.e., comment and comment thread). Moreover, it is rare to observe a learner repeatedly contributing to a particular parent post (i.e., comment thread). The analysis of the study data revealed 65% of learners who initiate a comment-thread have not posted any reply comments to other threads. Such discussion forum structures require examination of different aspects of role modelling, given the range of roles identified in the literature. Thus, to overcome the challenges in such MOOCs, this thesis classifies the learner roles as information-seeker, information-giver and other using the grounded theory approach.

This thesis only focuses on the linguistic features extracted from discussion forum data to promote real-time investigations. It is not always easy to incorporate contextual features

(e.g., votes, and views) in predictions, as they change throughout the course. Temporal changes in contextual features will not support the real-time role identification as contextual features for one single post can change throughout the course. For example, votes and views for one single post might change during the course. Moreover, psychological measures like cognitive skills, analytical ability can only be understood when investigating learners' discourse. In doing so, this thesis uses NLP tools and techniques to extract linguistic features from discussion forum posts to automate the content analysis. These tools and techniques have been proven to be an efficient way to analyse the human mind through different psychological measures. This thesis used the LIWC tool, a widely-used tool in linguistic research for content analysis. One of the key measures of this thesis is to investigate students' learning through linguistic features. The LIWC tool has the ability to demonstrate a person's psychological process through different measures, such as attentional focus and thinking styles. The tool extracts a wide range of linguistic features, such as summary dimensions (e.g., word counts, words per sentence, and analytic skills), and cognitive processes (e.g., differentiation) and has demonstrated its ability to detect learner roles and explain linguistic behaviours of different grades of learners. Topic models were also utilised to identify topics discussed in learner posts in this thesis. These NLP tools and techniques can be integrated into a real MOOC environment to automate the content analysis of the discussion forum posts and achieve the goal of this thesis.

Using the roles and linguistic features, the goal of this thesis was approached by exploring three main research studies, detailed in Chapters 5-7 of this thesis. Chapter 5 presented role prediction in learner posts. Chapter 6 discussed topic identification along with visualisation across different learner clusters, including the roles identified in Chapter 5. Finally, Chapter 7 presented a study on the relationship between linguistics and learner grades.

This chapter discusses the thesis contributions, along with the practical implications of these different studies to investigate the learning. Furthermore, this chapter presents threats to validity, and future work.

8.1 Investigating Student Learning

This thesis investigates the student learning in MOOCs through three different studies, namely role modelling, topic modelling, and linguistic analysis, which were presented in Chapters 5, 6, and 7 respectively. Using these approaches, it is evident that exploring learner roles, learner topics and linguistic behaviours helps to investigate students' learning in a MOOC environment.

The study presented in Chapter 5 focused on three research questions relating to roles as follows: 'RQ1: To what degree of granularity can a machine learning model predict learner roles in discussion forum posts using linguistic features alone?' 'RQ2: What are the linguistic features that contribute significantly towards identifying a learner role that is demonstrated in a forum post?' 'RQ3: To what extent can machine learning models that rely on linguistic features be used across courses from similar domains?'. The linguistic analysis was initially focused on topic modelling (as presented in Chapter 6), which addresses the following research questions: 'RQ1: What are the main discussion topics discussed in learner posts?' and 'RQ2: What are the main discussion topics discussed in different learner clusters?'. Finally, in Chapter 7, the linguistic analysis was developed using the research questions: 'RQ1: How do linguistic features extracted from students' discussion forum posts contribute to learner grade predictions?', 'RQ2: What are the significant rules that can be developed using the linguistic features extracted from discussion forums to identify the likelihood of different learner grades?' and 'RQ3: What are the significant linguistic features that can contribute to developing linguistic profiles of learners?'.

The following section describes the major contribution of each study presented in this thesis and articulates how they can be used to investigate students' learning.

8.1.1 Role Modelling

Identifying 'learner roles' is an important element in examining students' learning. This thesis presented a multi-class learner role classification model using linguistic-only features with the intention of eliminating the drawbacks that exist in previous studies. The model was built using the data obtained from the 'Introduction to Project Management' course to predict learner roles in discussion forum posts. The predictive model

performed well compared with the baseline model (Hecking et al., 2017), with an 87% F-measure. Furthermore, this model has been validated with the data obtained from the 'Risk Management for Projects' course to ensure it can be used across courses from similar domains.

Using a comprehensive dataset of 8,300 student posts, this thesis annotated the data into three learner roles ('information-giver', 'information-seeker' and 'other'). Each post was annotated into one of the aforementioned roles that best describes the forum post. This annotated dataset is the first contribution of this thesis, enabling other researchers to train a predictive model using the annotated data and test them in their intended courses. Furthermore, this thesis has also presented the detailed methodology for annotating a dataset using the grounded theory approach to identify the learner roles that best describe them. This methodology can be applied to any discussion forum data to identify user roles; not only in MOOCs but also in other forum platforms.

With these roles identified, the next contribution of this thesis is a multi-class predictive model that was built using the linguistic features alone. These linguistic features were extracted from learners' discussion forum posts using the LIWC tool. Furthermore, feature engineering was performed to extract the optimal features to identify a learner role in a forum post. This thesis also performed hyper parameter tuning and presented the hyper parameters. These model parameters can be replicated while building a predictive model for learner role identification in any other courses.

This thesis has listed a set of significant linguistic features that can be used to identify learner roles. To accommodate the model into various courses, the model was validated with discussion forum posts from a similar course 'Risk Management for Projects' to ensure the re-usability of the model. This thesis also presented a detailed methodology for role modelling whereby dissimilar courses can build predictive models and apply them to their courses. This role identification process can be implemented by training a model on the past years' course discussion forum data and can be used for predicting the roles in subsequent course offerings.

This thesis has presented a multi-class predictive model for role prediction that can be integrated into any real-world MOOC learning environment that is similar to the course used in this thesis. Roles can be used in many ways to investigate student learning. Firstly, they can be used as a filtering parameter in the course discussion forums such

that instructors can easily filter the roles of 'information-giver', 'information-seeker' and 'other'. Furthermore, this thesis has used only those discourse features that help in real-time role predictions. Therefore, it is an easy process for instructors to filter them promptly, based on the roles, and give feedback for urgent posts on time.

It is obvious that a learner who is an information-giver at a given time-frame does not have any difficulties and understand the course topic to some extent as they give information, while a learner who seeks information has some difficulties with understanding the course content. A learner who continuously identifies as an information-giver can be given less priority in terms of instructor interventions. However, the validity of their forum information needs to be examined. This can be validated using the votes given for the particular posts, as mentioned in the previous literature (J.-S. Wong et al., 2015). If an information-giver obtains high up votes for their posts, there is a high probability that they do not need any further guidance from the instructors on the given discussion topic. Conversely, if an information-giving learner continuously obtains negative votes for their posts, it is an indication that they did not understand the particular course topic and they are giving false information to their peers. This situation needs to be addressed by the instructors on time as they not only misunderstood the learning topic but are also misleading their peers.

On the other hand, information-seeking posts need to be addressed by instructors as a time-critical matter. These seeking posts emerged for a range of reasons, such as students being confused about a learning topic or wishing to ask subsequent questions to clarify their doubts. Studies such as (Yang et al., 2015; Agrawal et al., 2015; Z. Zeng et al., 2017; Atapattu et al., 2019) have also identified, cues (e.g., questions marks) that are used to identify seeking behaviour and have been identified as high predictors of confused learners. Furthermore, a learner who continuously seeks information needs to be given attention throughout the course as there is a high chance of attrition in such situations.

The analysis demonstrates learner roles can be used as an initial indicator to investigate students' learning in MOOCs. On the other hand, it is also important to identify the topics that are repeatedly questioned by many learners in order to provide generalised interventions in situations where giving personalised interventions requires a time-scale that is impossible due to the sheer numbers involved in the course. Therefore, the second study in this thesis focuses on topics that are discussed by learners and visualises them

across different learning clusters to investigate students' learning success across different learning areas.

8.1.2 Topic Modelling

The second study presented in Chapter 6 identified learner topics in the discussion forums and explored the relationship between these topics and different learner clusters, along with the learner roles that were identified in Chapter 5.

This thesis identified a set of topics using the Latent Dirichlet Allocation (LDA) topic model. The model was trained using the lecture materials along with an extended training corpus that consisted of similar Wikipedia content and learner posts from a different semester. The topic model extracted 13 topics along with the key terms that best describe them.

The next contribution of this research is the trained LDA topic model for the 'Project Management' course that was trained on the extended corpus using the seeding method (Cai et al., 2018). By applying the trained LDA model, the topics that were discussed in learner posts were predicted. Subsequently, this thesis explored the topics across different learner clusters, such as information-givers, information-seekers, High Distinction information-givers, and Fail-grade information-seekers.

In doing so, this thesis visualised the topic distribution across the learner roles that were identified in Chapter 5. Since the course is guided by a set of questions from the instructors, it was identified that information-givers have a higher percentage of topic contributions than information-seekers. This thesis further expanded the analysis and identified several interesting observations. It was identified that the number of learning activities and lecture materials affects the contributions of learners. A wide range of topics delivered under a certain module can make a student to become an information-seeker.

Moreover, the analysis discovered introducing case studies that can illustrate real-world scenarios can help the learners to engage more in the discussion forums. This finding can be used as a strategy by instructors so that they use real-world scenarios to explain a concept in a course. These real-world scenarios can be easily understood by learners, as the findings show information-seekers contribute less towards these case studies.

Furthermore, learners tend to share their knowledge by interpreting these real-world scenarios. Another finding of this thesis that can be used as a strategy to increase learners' engagement, is to ask learners to share their own experiences related to a course topic. It is always important to encourage students' forum participation, as instructors can measure students' learning through their discourse during the course and therefore do not need to wait till the end when adaptations can no longer be made.

Another interesting finding of this study reveals that the major contributions of information-seekers were identified towards the end of course. This shows the majority of learners seek to clarify their issues with the course modules that were delivered towards the end of the course. This finding can be implemented in a real MOOC environment for the subsequent offering of this course where additional instructors or student monitors can be appointed towards the end to rapidly clarify the issues before final assessments are completed. This might help learners to obtain a better grade in final assessments.

The analysis also highlights a descriptive course topic, where a topic is delivered using several sub-topics, can provoke learners' information-seeking behaviour. A similar trend is also observed for information-givers. Therefore, instructors need to be aware that introducing several new topics in a course module can increase learners' participation and there should be enough instructors during that time to provide feedback without delay.

Further analysis of learner roles across different course grade reveals Fail-grade learners give more information than Pass-grade and Credit-grade learners. Such information-giving posts need to be validated by the instructors, as they can mislead their peers. Furthermore, Fail-grade learners tend to seek information more often than other grades except High Distinction. This shows that learner engagement does not always positively correlate with learner grades, as learner participation can be affected by other factors, such as mandatory participation and the confusion state of a learner.

The results of this analysis show the topics that are contributed to by different learner groups. These findings can be implemented in course design of any subsequent offering or used as strategies by instructors in an on-going course. A course topic that has high information-seekers can be modified in subsequent offerings to help students to understand it more effectively. A course can be modified, for example, by incorporating real-world case studies to explain the topics that were previously less well understood.

Furthermore, constant instructor interventions can be provided more towards the end of the course, as these information-seekers are visible towards the end of the course.

This study shows that not only identifying topics but also visualising them, along with different learner clusters (e.g. topic distribution for information seekers - Figure 6.12) will help instructors to identify the topics that need immediate attention from learners; maybe through instructor interventions or by introducing brief case studies during the course to explain the topics that were not understood by learners. Furthermore, topics that have high numbers of information-givers can be updated with additional reading materials during the course. This helps to broaden their knowledge on the topic as they have already understood the fundamentals delivered in the course, further enhancing overall learning. In a real-world MOOC platform, a dashboard that integrates these features (e.g., roles versus topics) may help instructors in real-time with easy visualisations.

Apart from visualising the topics, it is also important to identify learners who are likely to obtain Pass-grades and Fail-grades in the final assessments and explore their discourse features. Therefore, the next study in this thesis was conducted to identify the correlations between learners' linguistic behaviours and learner grades.

8.1.3 Linguistic Analysis

The third study presented in Chapter 7 aimed to investigate the contribution of linguistic aspects that are extracted from learners' posts in their final course grades. In doing so, this thesis presented a set of rules to the research community to explain the relationship between the language features and learner grades. These decision rules were extracted from a prediction model that uses the decision tree algorithm. This prediction model was built using the linguistic features extracted from the LIWC tool, where the features were calculated in accordance with the weekly contributions.

This thesis contributes towards a set of 25 rules for the MOOC and 'Learning Analytics' research communities that can predict grades with more than 80% probability. These rules are presented along with the probability of predicting a learner grade using the learners' linguistic behaviour in discussion forums. Though these rule sets include general linguistic features (i.e. not course specific linguistic features), these rule sets can be initially applied to the subsequent course offerings to ensure its validity. Further,

these rule sets can be modified for dissimilar courses by applying similar rule extraction methodology that was described in section 7.2.

According to the analysis, the level of 'Analytical thinking' expressed during the 5^{th} week of the course has the highest information gain to predict learner grades. However, the thesis also presents a set of linguistic features that can be identified at the beginning of the course, such as use of 'personal pronoun ('I')', 'negative emotions', and 'comparisons'. Therefore, investigating such attributes presented at the early stage of a course can be used for predicting the likelihood of learners' grade outcomes. These linguistic behaviours help to identify learners who are likely to fail in the course so that instructors can provide sufficient support throughout the course to enhance their learning.

It is found that learners' expressing lower levels of analytical ability during the end of the course are more likely to obtain either Pass or Fail course grades. However, the results emphasise investigating the values of 'Compare' expressed at week 3 of the course helps to identify the learners' grade, where the majority of the Fail-grade learners use fewer comparison language terms. The results also demonstrate that higher personal pronoun ('I') usage at week 2 will always result in a Fail-grade with more than 80% probability. This is an interesting linguistic cue that reveals that those learners' who express personal pronouns often require more attention than other learners, as they tend to fail the course.

These rule sets demonstrate investigating the continuous contributions of a learner in discussion forums helps to understand their learning. Furthermore, these rule sets are applicable even if a contribution from a learner is not observed for a few weeks. In a real-world MOOC environment, along with optional participation, it is hard to expect that learners will continuously post in forums; however, these rule sets can be applied to learners who do not post continuously. These rule sets can be applied in a real MOOC environment to determine the likelihood of final course grades by observing the values for the linguistic features that have the highest information gain. They give clues to instructors about where students are at in the learning life-cycle and whether they need instructor interventions.

Finally, this thesis visualised how these linguistic features change during the course by examining 'Comment Thread' and 'Comments' of the discussion forum, as presented in Figure 4.1. These analyses are useful when a learner posts in discussion forums. Statistical analysis reveals the linguistic features that significantly define both Pass-grade

and Fail-grade linguistic profiles. According to the Pass-grade linguistic profile, linguistic features such as Analytic, Words per Sentence, Differentiation (i.e., Cognitive Processes) will be high in their discourse. Conversely, a Fail-grade linguistic profile shows they hold high values for linguistic features such as Clout, Dictionary words, Personal Pronouns, and informal language. These distinctive linguistic profiles can be used as a template for instructors to track students and understand their level of learning at the given time (i.e., whether they obtain Pass or Fail based on their contributions on discussion forum).

8.2 Threats to Validity

There are several threats to validity for this research. The main limitation of this thesis is that these findings cannot be applied to learners who do not entirely participate in the discussion forums. Discussion forums that do not require mandatory participation result in several learners who do not contribute to forum discussions. Such learners' learning cannot be investigated through the measures (i.e., roles and linguistic expressions) presented in this thesis.

Learners' language competency is the next limitation that can affect the results of this investigation. Those learners who are native English speakers tend to answer the questions in a descriptive way, whilst learners who speak English as their second language (or more) tend to provide an informative answer using a more limited vocabulary. These types of discourse can affect the prediction model and can easily be misinterpreted by the models.

Another threat to validity is the predictive models that were presented in this thesis were not tested on dissimilar courses. However, courses from dissimilar domains (e.g., Medicine) can replicate the methodologies that were presented in this thesis. They can be replicated by training a new model on a course's past data, as these models need to be trained with the existing forum data to be more accurate.

Another limitation is the inability to capture new topics using the pre-trained topic model. The topic model used in this thesis was trained with the existing discussion forum data. The model is unable to capture new interesting learning topics that were not presented in the previous offerings. However, the training corpus can be expanded

using the seeding methodology that was presented in this thesis and retraining the model with the new corpus enables the model to capture entirely new topics.

Another limitation of the thesis is that these findings can be influenced by the learners' personalities and their learning context. Every year, new learners with different personalities join courses: their study context may be entirely different from earlier cohorts. Therefore, generalising this model to each and every learner and expecting the same results may not be possible as the success of their learning may depend on other factors.

8.3 Future Work

This section presents future research directions arising from the work presented in this thesis. It discusses the future research opportunities for each study conducted in this thesis.

- A study evaluating the predictive model on roles in dissimilar courses.

  This thesis has contributed an annotated data set to the learning community such that researchers could use this as training data and test the data on another course. This thesis has already validated the model in a course from a similar domain and obtained an 86% F-measure. It would be interesting to investigate further how this predictive model would perform in courses from different domains such as Medicine and Computer Science. Although this study has contributed towards a predictive model that uses domain-independent features, it would be interesting to evaluate its performance on entirely dissimilar courses.

- A study examining changing linguistic patterns while switching from one role to another.

  This thesis presented a predictive model to predict roles expressed in discussion forums. During a course, a learner can express different roles in their distinctive forum posts. Interesting conclusions can be made from a study that analyses the linguistic differences when switching from one role to another (e.g., IS →IG or IG →IS or O →IG) versus expressing the same role in consecutive posts (e.g., IS →IS or IG →IG). For example, examining how the cognitive process changes when a learner switches from information-giver →information-seeker. Similarly, it would

be helpful to examine the cognitive changes that occur when a learner expresses the same role in consecutive posts.

- Explore the identified discussion topics with time.

  This thesis has contributed to a topic model, visualises the topics across different leaner groups and highlights the findings that can be implemented as strategies in a MOOC environment. During the learning process, a learners' interest can change over time as the course progresses. Hence, it would be interesting to explore how the learner interest (i.e., discussion topics) change with time. Furthermore, investigations into how a topic contribution time frame changes between different learner groups especially, Fail-grade and High Distinction learners, and identification of any interesting observations that may help instructors to derive new learning strategies would provide better support to both learners and instructors.

- Analysing the relationship between comment threads and comments in terms of topics and learner roles.

  It would be interesting to investigate the relationships between comment threads and comments, as it may provide useful findings to MOOC instructors. For example, analysing the contribution of information-giving comments for an information-seeking thread. This shows, how much an information-seeking question being addressed by learners' comments or whether there are any further information-seeking comments being observed.

- Implement the grade prediction model with progressively more data.

  The thesis has implemented the grade prediction model with entire discussion forum data (i.e. From Week 1 to Week 6). However, it is important to identify the right amount of weekly discussion forum data to enable the early identification of at-risk students. This can be achieved by developing a series of predictive models with progressively more data. For example, the initial predictive model can be built using only the data from the first two weeks and the subsequent model can be built by adding data from the third week. This will help to determine the amount of weekly data that is required to train a model that is trustworthy enough to serve as the basis for identifying at-risk learners early in the course.

- Generating a prediction model to forecast linguistic features.

This thesis has visualised the linguistics with time in comment threads and comments. A time series prediction model to predict learners' next set of linguistic features in coming weeks and analysing their trend would be a next step. An in-depth study to predict the linguistic changes in comment threads and comments would be beneficial for long term courses that span across a year.

- Investigate 'Linguistic Profiles' for each week

  This thesis presented Linguistic Profiles for Pass-grade and Fail-grade learners that were curated by extracting the significant differences between them. However, this thesis has also visualised these features with time and identified there is a divergence in learners' language during the course. Hence, exploring the linguistic profile each week could help to identify and expand our understanding of a learners' linguistic patterns.

- Derive a set of rules from decision tree classifiers to identify the distinctive linguistic behavior for further distinctive categories by incorporating user roles (information-givers and information-seekers).

  The thesis has derived rule sets to identify the linguistic behavior of pass and fail grade learners. Similar analysis can be replicated to identify the distinctive linguistic behavior for the learner roles combined along with learner grades. Such analysis can provide more fine-grained insights.

8.4 Concluding Remarks

MOOC learning environments are gaining massively in popularity and attract a huge number of learners. While learners enroll with different purposes, completion is not the only indicator of learning success. Therefore, it is important to use other measures to investigate student learning in MOOC environments.

This thesis began with the intention to investigate students' learning through roles and linguistic expressions that were extracted from the learners' discourse in discussion forum posts. This thesis presented several predictive models and a topic model that can help to identify learner roles and discussion topics. Furthermore, significant linguistic rule sets and linguistic profiles were extracted and presented in this thesis. By building a series of

models using the discussion forum data, the thesis demonstrates to the learning analytics research community that it is possible to investigate students' learning by investigating the outcomes that are generated from these models.

Furthermore, this thesis also presented insights about student learning through the studies conducted on roles, learner topics across different learner clusters, and rules sets that reveal the relationship between linguistics and learner grades, particularly considering nature of optional participation. Strategies were also identified based on the research findings to help instructors provide a better learning environment for learners.

The thesis examines the students' learning through three studies using the discussion forum posts extracted from the AdelaideX learning platform. The thesis provides an annotated dataset that can be used by researchers for training and evaluating their machine learning models to identify learner roles. The thesis also presented the methodology to identify the learner roles in new discussion forum data and explains the detailed process of their annotation. This thesis demonstrates roles and linguistic expressions extracted from learner discourse have the potential to analyse students' learning and help to identify predicted learner success and potentially at-risk students. Furthermore, the methodologies involved in this thesis can be replicated across different courses. These methodologies serve as a guide for evaluating an entirely new MOOC discussion forum in the future.

It is expected that the outcomes of this thesis would be helpful for educators and researchers who are involved in the education industry, who seek to use trained machine learning models and the trained topic model to predict learner roles and analyse discussion topics. Furthermore, they will be able to apply the rule sets and significant language cues with high information gain; to identify the likelihood of learners' final course grades. The outcomes of this thesis could be used in designing and evaluating a wide range of MOOCs, or large online courses, enabling instructors to impose the identified strategies for learner retention. Finally, the contribution of this thesis holds great promise for further advancements in understanding student learning in MOOCs. It also helps in the development of new MOOC tools and platforms that can investigate students' learning in online learning platforms and supports advancement in learning analytics research within these environments.

References

Adaji, I., & Olakanmi, O. A. (2019). Evolution of emotions and sentiments in an online learning community. In *Slll@ aied* (pp. 23–27).

Agrawal, A., Venkatraman, J., Leonard, S., & Paepcke, A. (2015). Youedu: addressing confusion in mooc discussion forums by recommending instructional video clips.

Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., & Goodrich, V. (2014). Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 103–112).

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, *3*, 42.

Alexander-Emery, S., Cohen, L. M., & Prensky, E. H. (2005). Linguistic analysis of college aged smokers and never smokers. *Journal of Psychopathology and Behavioral Assessment*, *27*(1), 11–16.

Aljohani, N. R., Fayoumi, A., & Hassan, S.-U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability*, *11*(24), 7238.

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., & Taylor, C. B. (2005). Evaluation of computerized text analysis in an internet breast cancer support group. *Computers in Human Behavior*, *21*(2), 361–376.

Alshabandar, R., Hussain, A., Keight, R., & Khan, W. (2020). Students performance prediction in online courses using machine learning algorithms. In *2020 international joint conference on neural networks (ijcnn)* (pp. 1–7).

Alvarez-Conrad, J., Zoellner, L. A., & Foa, E. B. (2001). Linguistic predictors of trauma pathology and physical health. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *15*(7), S159–S170.

Alvero, A., Giebel, S., Gebre-Medhin, B., Antonio, A. L., Stevens, M. L., & Domingue, B. W. (2021). Essay content and style are strongly related to household income and sat scores: Evidence from 60,000 undergraduate applications. *Science Advances*, *7*(42), eabi9031.

Anderson, T. (2008). *The theory and practice of online learning.* Athabasca University Press.

Andy, A., Andy, U., et al. (2021). Understanding communication in an online cancer forum: Content analysis study. *JMIR cancer*, *7*(3), e29555.

Apaza, R. G., Cervantes, E. V., Quispe, L. C., & Luna, J. O. (2014). Online courses recommendation based on lda. In *Simbig* (pp. 42–48).

Arguello, J., Butler, B. S., Joyce, E., Kraut, R., Ling, K. S., Rosé, C., & Wang, X. (2006). Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 959–968).

Arguello, J., & Shaffer, K. (2015). Predicting speech acts in mooc forum posts. In *Ninth international aaai conference on web and social media.*

Ashenafi, M. M., Riccardi, G., & Ronchetti, M. (2015). Predicting students' final exam scores from their course activities. In *2015 ieee frontiers in education conference (fie)* (pp. 1–9).

Atapattu, T., & Falkner, K. (2016). A framework for topic generation and labeling from mooc discussions. In *Proceedings of the third (2016) acm conference on learning@ scale* (pp. 201–204).

Atapattu, T., Falkner, K., Thilakaratne, M., Sivaneasharajah, L., & Jayashanka, R. (2019). An identification of learners' confusion through language and discourse analysis. *arXiv preprint arXiv:1903.03286*.

Atapattu, T., Falkner, K., Thilakaratne, M., Sivaneasharajah, L., & Jayashanka, R. (2020). What do linguistic expressions tell us about learners' confusion? a domain-independent analysis in moocs. *IEEE Transactions on Learning Technologies*, *13*(4), 878–888.

Austin, J. L. (1962). How to do things with words. clarendon. *Oxford*, *2005*, 619–650.

Baddeley, J. L., & Singer, J. A. (2008). Telling losses: Personality correlates and functions of bereavement narratives. *Journal of Research in Personality*, *42*(2), 421–438.

Baghaei, N., Mitrovic, A., & Irwin, W. (2007). Supporting collaborative learning and problem-solving in a constraint-based cscl environment for uml class diagrams.

International Journal of Computer-Supported Collaborative Learning, *2*(2), 159–190.

Bantum, E. O., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological assessment*, *21*(1), 79.

Barab, S. A., & Duffy, T. (2000). From practice fields to communities of practice. *Theoretical foundations of learning environments*, *1*(1), 25–55.

Beaudreau, S. A., Storandt, M., & Strube, M. J. (2005). A comparison of narratives told by younger and older adults. *Experimental Aging Research*, *32*(1), 105–117.

Beevers, C. G., & Scott, W. D. (2001). Ignorance may be bliss, but thought suppression promotes superficial cognitive processing. *Journal of Research in Personality*, *35*(4), 546–553.

Belbin, R. M. (2010). *Management teams*. Routledge.

Bento, R., Brownstein, B., Kemery, E., Zacur, S. R., et al. (2005). A taxonomy of participation in online courses. *Journal of College Teaching & Learning (TLC)*, *2*(12).

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning* (pp. 115–123).

Beukeboom, C. (2014). Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. In *Social cognition and communication* (pp. 313–330). Psychology Press.

Beukeboom, C. J., Forgas, J., Vincze, O., & Laszlo, J. (2014). Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication*, *31*, 313–330.

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

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature*, *549*(7671), 195–202.

Biddle, B. J. (1986). Recent developments in role theory. *Annual review of sociology*, *12*(1), 67–92.

Biddle, B. J. (2013). *Role theory: Expectations, identities, and behaviors*. Academic Press.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Boals, A., & Klein, K. (2005). Word use in emotional narratives about failed romantic relationships and subsequent mental health. *Journal of Language and Social Psychology*, *24*(3), 252–268.

Bonafini, F., Chae, C., Park, E., & Jablokow, K. (2017). How much does student engagement with videos and forums in a mooc affect their achievement? *Online Learning Journal*, *21*(4).

Boroujeni, M. S., Hecking, T., Hoppe, H. U., & Dillenbourg, P. (2017). Dynamics of mooc discussion forums. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 128–137).

Buenaño-Fernandez, D., González, M., Gil, D., & Luján-Mora, S. (2020). Text mining of open-ended questions in self-assessment of university teachers: An lda topic modeling approach. *IEEE Access*, *8*, 35318–35330.

Burch, M., Lohmann, S., Pompe, D., & Weiskopf, D. (2013). Prefix tag clouds. In *2013 17th international conference on information visualisation* (pp. 45–50).

Burt, R. S. (1982). *Toward a structural theory of action: Network models of social structure, perception, and action*. academic Press.

Cai, Z., Graesser, A. C., Windsor, L., Cheng, Q., Shaffer, D. W., & Hu, X. (2018). Impact of corpus size and dimensionality of lsa spaces from wikipedia articles on autotutor answer evaluation. *Journal of educational data mining*.

Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, *26*(7), 675–693.

Chen, B., & Poquet, O. (2020). Socio-temporal dynamics in peer interaction events. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 203–208).

Cheng, C. K., Paré, D. E., Collimore, L.-M., & Joordens, S. (2011). Assessing the effectiveness of a voluntary online discussion forum on improving students' course performance. *Computers & Education*, *56*(1), 253–261.

Chi, M. T., & Wylie, R. (2014). The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, *49*(4), 219–243.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Clow, D. (2013). Moocs and the funnel of participation. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 185–189).

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of educational research*, *64*(1), 1–35.

Cohen, W., Carvalho, V., & Mitchell, T. (2004). Learning to classify email into "speech acts". In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 309–316).

Core, M. G., Moore, J. D., & Zinn, C. (2003). The role of initiative in tutorial dialogue. In *10th conference of the european chapter of the association for computational linguistics*.

Crossley, S., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. (2015). Language to completion: Success in an educational data mining massive open online class. *International Educational Data Mining Society*.

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with nlp tools to better understand mooc completion. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 6–14).

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes*, *51*(5-6), 511–534.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, *48*(4), 1227–1237.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, *49*(3), 803–821.

Dale, R. (1991). Exploring the role of punctuation in the signalling of discourse structure. In *Proceedings of a workshop on text representation and domain modelling: ideas from linguistics and ai* (pp. 110–120).

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words! linguistic style accommodation in social media. In *Proceedings of the 20th international conference on world wide web* (pp. 745–754).

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on world wide web* (pp. 307–318).

Darling-Hammond, L., Austin, K., Lit, I., & Nasir, N. (2003). The learning classroom: Theory into practice. In *Session 11—lessons for life: Learning and transfer.* Stanford University VT.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Devi, R., & Nirmala, K. (2013). Construction of decision tree: Attribute selection measures. *International Journal of Advancements in Research & Technology*, *2*(4), 343–347.

De Wever, B., Van Keer, H., Schellens, T., & Valcke, M. (2010). Roles as a structuring tool in online discussion groups: The differential impact of different roles on social knowledge construction. *Computers in Human Behavior*, *26*(4), 516–523.

Dillahunt, T., Wang, Z., & Teasley, S. D. (2014). Democratizing higher education: Exploring mooc use among those who cannot afford a formal education. *International Review of Research in Open and Distributed Learning*, *15*(5), 177–196.

D'Mello, S. K., & Graesser, A. (2012). Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*, *5*(4), 304–317.

Dowell, N. M., Brooks, C., Kovanović, V., Joksimović, S., & Gašević, D. (2017). The changing patterns of mooc discourse. In *Proceedings of the fourth (2017) acm conference on learning@ scale* (pp. 283–286).

Dowell, N. M., & Poquet, O. (2021). Scip: Combining group communication and interpersonal positioning to identify emergent roles in scaled digital environments. *Computers in Human Behavior*, *119*, 106709.

Dowell, N. M., Skrypnyk, O., Joksimovic, S., Graesser, A. C., Dawson, S., Gaševic, D., ... Kovanovic, V. (2015). Modeling learners' social centrality and performance through language and discourse. *International Educational Data Mining Society*.

Duncan, S. Y., Chohan, R., & Ferreira, J. J. (2019). What makes the difference? employee social media brand engagement. *Journal of Business & Industrial Marketing*.

Dupuy, C., Bach, F., & Diot, C. (2017). Qualitative and descriptive topic extraction from movie reviews using lda. In *International conference on machine learning and data mining in pattern recognition* (pp. 91–106).

Elkins, A., Freitas, F. F., & Sanz, V. (2019). Developing an app to interpret chest x-rays to support the diagnosis of respiratory pathology with artificial intelligence. *arXiv preprint arXiv:1906.11282*.

Evans, C. M. (2020). Measuring student success skills: A review of the literature on collaboration. 21st century success skills. *National Center for the Improvement of Educational Assessment*.

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding mooc discussion forums: a learning analytics approach. In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 146–150).

Ferguson, R., & Sharples, M. (2014). Innovative pedagogy at massive scale: teaching and learning in moocs. In *European conference on technology enhanced learning* (pp. 98–111).

Figueiredo, S., Soares, A., Vieira, N., Devezas, M., et al. (2020). A psycholinguistic analysis of world leaders? discourses concerning the covid-19 context: Authenticity and emotional tone. *International Journal of Social Sciences*, *9*(2), 66–69.

Flores, F., & Ludlow, J. J. (1980). Doing and speaking in the office. *Decision support systems: Issues and challenges*, 95–118.

Friedman, S. R., Rapport, L. J., Lumley, M., Tzelepis, A., VanVoorhis, A., Stettner, L., & Kakaati, L. (2003). Aspects of social and emotional competence in adult attention-deficit/hyperactivity disorder. *Neuropsychology*, *17*(1), 50.

Gaebel, M. (2014). *Moocs: Massive open online courses* (Vol. 11). EUA Geneva.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68–84.

Geddes, D. (1988). Linguistic markers of the organizational/group identification process..

Giles, H., Coupland, N., & Coupland, I. (1991). 1. accommodation theory: Communication, context, and. *Contexts of accommodation: Developments in applied sociolinguistics*, *1*.

Gill, A. J., French, R. M., Gergle, D., & Oberlander, J. (2008). The language of emotion in short blog texts. In *Proceedings of the 2008 acm conference on computer supported cooperative work* (pp. 299–302).

Glaser, B. G. (1978). Theoretical sensitivity. mill valley.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, *37*(1), 3–19.

Gortner, E.-M., & Pennebaker, J. W. (2003). The archival anatomy of a disaster: Media coverage and community-wide health effects of the texas a&m bonfire tragedy. *Journal of Social and Clinical Psychology*, *22*(5), 580–603.

Gottschalk, L. A., & Gleser, G. C. (1979). *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193–202.

Guastella, A. J., & Dadds, M. R. (2006). Cognitive-behavioral models of emotional writing: A validation study. *Cognitive Therapy and Research*, *30*(3), 397–414.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Gütl, C., Rizzardini, R. H., Chang, V., & Morales, M. (2014). Attrition in mooc: Lessons learned from drop-out students. In *International workshop on learning technology for education in cloud* (pp. 37–48).

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), 389–422.

Hämäläinen, W., & Vinni, M. (2011). Classifiers for educational data mining. *Handbook of Educational Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, 57–71.

Hare, A. P. (1994). Types of roles in small groups: A bit of history and a current perspective. *Small Group Research*, *25*(3), 433–448.

Hartley, J., Pennebaker, J., & Fox, C. (2003). Abstracts, introductions and discussions: How far do they differ in style? *Scientometrics*, *57*(3), 389–398.

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

Hecking, T., Chounta, I.-A., & Hoppe, H. U. (2017). Role modelling in mooc discussion forums. *Journal of Learning Analytics*, *4*(1), 85–116.

Ho, A., Reich, J., Nesterko, S., Seaton, D., Mullaney, T., Waldo, J., & Chuang, I. (2014). Harvardx and mitx: The first year of open online courses, fall 2012-summer 2013. *Ho, AD, Reich, J., Nesterko, S., Seaton, DT, Mullaney, T., Waldo, J., & Chuang, I.(2014). HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1)*.

Ho, I. M. K., Cheong, K. Y., & Weldon, A. (2021). Predicting student satisfaction of emergency remote learning in higher education during covid-19 using machine learning techniques. *Plos one*, *16*(4), e0249423.

Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent dirichlet allocation. *advances in neural information processing systems*, *23*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international acm sigir conference on research and development in information retrieval* (pp. 50–57).

Hollands, F. M., & Tirthali, D. (2014). Resource requirements and costs of developing and delivering moocs. *International Review of Research in Open and Distributed Learning*, *15*(5), 113–133.

Hu, J., Dowell, N., Brooks, C., & Yan, W. (2018). Temporal changes in affiliation and emotion in mooc discussion forum discourse. In *International conference on artificial intelligence in education* (pp. 145–149).

Hu, X., Zhang, X., Lu, C., Park, E. K., & Zhou, X. (2009). Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 389–396).

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behavior in mooc forums. In *Proceedings of the first acm conference on learning@ scale conference* (pp. 117–126).

Huffaker, D., Jorgensen, J., Iacobelli, F., Tepper, P., & Cassell, J. (2006). Computational measures for language similarity across time in online communities. In *Proceedings of the analyzing conversations in text and speech* (pp. 15–22).

Jagarlamudi, J., Daumé III, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. In *Proceedings of the 13th conference of the european chapter of the association for computational linguistics* (pp. 204–213).

Jarboe, S. (1996). *Procedures for enhancing group decision making in hirokawa, b. & poole, m.(eds) communication and group decision making.* Thousand Oaks, Ca: Sage Publications.

Jeong, H., Hmelo-Silver, C. E., & Jo, K. (2019). Ten years of computer-supported collaborative learning: A meta-analysis of cscl in stem education during 2005–2014. *Educational research review*, *28*, 100284.

Jiang, S., Williams, A., Schenke, K., Warschauer, M., & O'dowd, D. (2014). Predicting mooc performance with week 1 behavior. In *Educational data mining 2014*.

Johnson, L., Becker, S., Estrada, V., & Freeman, A. (2014). *Nmc horizon report: 2014 higher education edition* (Tech. Rep.). Austin, Texas: The New Media Consortium. Retrieved from https://www.learntechlib.org/p/130341

Joksimovic, S., Baker, R. S., Ocumpaugh, J., Andres, J. M. L., Tot, I., Wang, E. Y., & Dawson, S. (2019). Automated identification of verbally abusive behaviors in online discussions. In *Proceedings of the third workshop on abusive language online* (pp. 36–45).

Jones, B. (1995). Exploring the role of punctuation in parsing natural text. *arXiv preprint cmp-lg/9505024*.

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *International Review of Research in Open and Distributed Learning*, *15*(1), 133–160.

Jordan, K. N., Sterling, J., Pennebaker, J. W., & Boyd, R. L. (2019). Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences*, *116*(9), 3476–3481.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143.

Kim, M. K., Wang, Y., & Ketenci, T. (2020). Who are online learning leaders? piloting a leader identification method (lim). *Computers in Human Behavior*, *105*, 106205.

Kim, S. N., Wang, L., & Baldwin, T. (2010). Tagging and linking web forum posts. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 192–202).

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of*

the third international conference on learning analytics and knowledge (pp. 170–179).

Klare, G. R. (1974). Assessing readability. *Reading research quarterly*, 62–102.

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting mooc dropout over weeks using machine learning methods. In *Proceedings of the emnlp 2014 workshop on analysis of large scale social interaction in moocs* (pp. 60–65).

Kollar, I., Fischer, F., & Hesse, F. W. (2006). Collaboration scripts–a conceptual analysis. *Educational Psychology Review*, *18*(2), 159–185.

Koller, D., Ng, A., Do, C., & Chen, Z. (2013). Retention and intention in massive open online courses: In depth. *Educause review*, *48*(3), 62–63.

Koschmann, T. (1996). Paradigm shifts and instructional technology: An introduction. *CSCL: Theory and practice of an emerging paradigm*, *116*, 1–23.

Kovanović, V., Joksimović, S., Gašević, D., Siemens, G., & Hatala, M. (2015). What public media reveals about mooc s: A systematic analysis of news reports. *British Journal of Educational Technology*, *46*(3), 510–527.

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018). Understand students' self-reflections through learning analytics. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 389–398).

Kowalski, R. M. (2000). "i was only kidding!": Victims' and perpetrators' perceptions of teasing. *Personality and Social Psychology Bulletin*, *26*(2), 231–241.

Kuang, W., Luo, N., & Sun, Z. (2011). Resource recommendation based on topic model for educational system. In *2011 6th ieee joint international information technology and artificial intelligence conference* (Vol. 2, pp. 370–374).

Kuang, X., Chae, H., Hughes, B., & Natriello, G. (2021). An lda topic model and social network analysis of a school blogging platform. In *Proc. 10th international conference on educational data mining* (pp. 362–363).

Kuh, G. D. (2003). What we're learning about student engagement from nsse: Benchmarks for effective educational practices. *Change: The magazine of higher learning*, *35*(2), 24–32.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, *49*(4), 757–786.

Lewin, K. (1942). Time perspective and morale.

Li, X., Xie, L., & Wang, H. (2016). Grade prediction in moocs. In *2016 ieee intl conference on computational science and engineering (cse) and ieee intl conference on embedded and ubiquitous computing (euc) and 15th intl symposium on distributed computing and applications for business engineering (dcabes)* (pp. 386–392).

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: A big comparison for nas. *arXiv preprint arXiv:1912.06059*.

Litvinova, T., & Litvinova, O. (2018). A study of texts of an extremist forum "kavkazchat" using linguistic inquiry and word count (liwc). *QUALICO*, *2018*, 69.

Liu, W., Kidziński, Ł., & Dillenbourg, P. (2016). Semiautomatic annotation of mooc forum posts. In *State-of-the-art and future directions of smart learning* (pp. 399–408). Springer.

Löckenhoff, C. E., Costa Jr, P. T., & Lane, R. D. (2008). Age differences in descriptions of emotional experiences in oneself and others. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *63*(2), P92–P99.

Lohmann, S., Ziegler, J., & Tetzlaff, L. (2009). Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Ifip conference on human-computer interaction* (pp. 392–404).

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

Mahmood, R. K. (2019). The dissolution of linguistics and the rise of language with reference to pragmatics: A deconstructive approach. *Journal of University of Human Development*, *5*(3), 1–5.

Mahmood, R. K., et al. (2016). Pragmatics between microlinguistic and macrolinguistic levels of analysis. *Global Journal of Foreign Language Teaching*, *6*(3), 126–129.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1–15.

Marcos-García, J.-A., Martínez-Monés, A., & Dimitriadis, Y. (2015). Despro: A method based on roles to provide collaboration analysis support adapted to the participants in cscl situations. *Computers & Education*, *82*, 335–353.

Mayer, D. K., Terrin, N. C., Kreps, G. L., Menon, U., McCance, K., Parsons, S. K., & Mooney, K. H. (2007). Cancer survivors information seeking behaviors: a comparison of survivors who do and do not seek information about cancer. *Patient education and counseling*, *65*(3), 342–350.

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The mooc model for digital practice.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. *http://mallet. cs. umass. edu*.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods*, *45*(2), 499–515.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with coh-metrix*. Cambridge University Press.

Medina, E., Vega, D., Meseguer, R., Medina, H., Ochoa, S. F., & Magnani, M. (2016). Using indirect blockmodeling for monitoring students roles in collaborative learning networks. In *2016 ieee 20th international conference on computer supported cooperative work in design (cscwd)* (pp. 164–169).

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, *84*(4), 857.

Mey, G., & Mruck, K. (2011). Grounded-theory-methodologie: Entwicklung, stand, perspektiven. In *Grounded theory reader* (pp. 11–48). Springer.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272).

Moore, R. L., Yen, C.-J., & Powers, F. E. (2021). Exploring the relationship between clout and cognitive processing in mooc discussion forums. *British Journal of Educational Technology*, *52*(1), 482–497.

Mudrack, P. E., & Farrell, G. M. (1995). An examination of functional role behavior and its consequences for individuals in group settings. *Small Group Research*, *26*(4), 542–571.

Nanda, G., Douglas, K. A., Waller, D. R., Merzdorf, H. E., & Goldwasser, D. (2021). Analyzing large collections of open-ended feedback from mooc learners using lda topic modeling and qualitative analysis. *IEEE Transactions on Learning Technologies*.

Nanda, G., Hicks, N. M., Waller, D. R., Goldwasser, D., & Douglas, K. A. (2018). Understanding learners' opinion about participation certificates in online courses using topic modeling. *International Educational Data Mining Society*.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 100–108).

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, *29*(5), 665–675.

Nguyen, D., & Rose, C. (2011). Language use as a reflection of socialization in online communities. In *Proceedings of the workshop on language in social media (lsm 2011)* (pp. 76–85).

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360.

Niraula, N., Banjade, R., Ştefănescu, D., & Rus, V. (2013). Experiments with semantic similarity measures based on lda and lsa. In *International conference on statistical language and speech processing* (pp. 188–199).

O'dea, B., Larsen, M. E., Batterham, P. J., Calear, A. L., & Christensen, H. (2017). A linguistic analysis of suicide-related twitter posts. *Crisis*.

OECD. (2013). *Pisa 2015 collaborative problem solving framework*. OECD Publishing Paris, France.

Ofran, Y., Paltiel, O., Pelleg, D., Rowe, J. M., & Yom-Tov, E. (2012). Patterns of information-seeking for cancer on the internet: an analysis of real world data.

Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017). A neural network approach for students' performance prediction. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 598–599).

Oliver, K. M., Houchins, J. K., Moore, R. L., & Wang, C. (2021). Informing makerspace outcomes through a linguistic analysis of written and video-recorded project assessments. *International Journal of Science and Mathematics Education*, *19*(2), 333–354.

Onah, D., Sinclair, J., Boyatt, R., & Foss, J. (2014). Massive open online courses: Learner participation. In *Proceeding of the 7th international conference of education, research and innovation* (pp. 2348–2356).

Onah, D. F., & Pang, E. L. (2021). Mooc design principles: topic modelling-pyldavis visualization & summarisation of learners' engagement..

Öztuna, D., Elhan, A. H., & Tüccar, E. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*, *36*(3), 171–176.

Palmer, S., Holt, D., & Bray, S. (2008). Does the discussion help? the impact of a formally assessed online discussion on final student results. *British Journal of Educational Technology*, *39*(5), 847–858.

Pasupathi, M. (2007). Telling and the remembered self: Linguistic differences in memories for previously disclosed and previously undisclosed events. *Memory*, *15*(3), 258–270.

Patel, J., & Aghayere, A. (2006). Students' perspective on the impact of a web-based discussion forum on student learning. In *Proceedings. frontiers in education. 36th annual conference* (pp. 26–31).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825-2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

Pennebaker, J., Booth, R., Boyd, R., & Francis, M. (2015). Liwc 2015 operator's manual. *Austin, TX: Pennebaker Conglomerates Inc.*

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). Linguistic inquiry and word count: Liwc2015.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, *135*.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015* (Tech. Rep.).

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS one*, *9*(12), e115844.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of liwc2007: Liwc. net. *Google Scholar*.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, *10*(3), 90–93.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547–577.

Perna, L. W., Ruby, A., Boruch, R. F., Wang, N., Scull, J., Ahmad, S., & Evans, C. (2014). Moving through moocs: Understanding the progression of users in massive open online courses. *Educational Researcher*, *43*(9), 421–432.

Phillips, W. J. (2018). Past to future: Self-compassion can change our vision. *Journal of Positive Psychology and Wellbeing*.

Postmes, T., Spears, R., & Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human communication research*, *26*(3), 341–371.

Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Qadir, A., & Riloff, E. (2011). Classifying sentences as speech acts in message board posts. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 748–758).

Rajasundari, T., Subathra, P., & Kumar, P. (2017). Performance analysis of topic modeling algorithms for news articles. *Journal of Advanced Research in Dynamical and Control Systems*, *11*, 175–183.

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2013). Modeling learner engagement in moocs using probabilistic soft logic. In *Nips workshop on data driven education* (Vol. 21, p. 62).

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2014). Understanding mooc discussion forums using seeded lda. In *Proceedings of the ninth workshop on innovative use of nlp for building educational applications* (pp. 28–33).

Richhariya, B., Tanveer, M., Rashid, A., Initiative, A. D. N., et al. (2020). Diagnosis of alzheimer's disease using universum support vector machine based recursive feature elimination (usvm-rfe). *Biomedical Signal Processing and Control*, *59*, 101903.

Roberts, T. S. (2005). Computer-supported collaborative learning in higher education. In *Computer-supported collaborative learning in higher education* (pp. 1–18). IGI Global.

Robins, R. H. (2014). *General linguistics*. Routledge.

Robinson, A. C. (2015). Exploring class discussions from a massive open online course (mooc) on cartography. In *Modern trends in cartography* (pp. 173–182). Springer.

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016). Forecasting student achievement in moocs with natural language processing. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 383–387).

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368–384.

Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R., & VanLehn, K. (2003). Exploring the effectiveness of knowledge construction dialogues. *Artificial intelligence in education: Shaping the future of learning through intelligent technologies*, 497–499.

Rossetti, M., Stella, F., & Zanker, M. (2016). Analyzing user reviews in tourism with topic models. *Information Technology & Tourism*, *16*(1), 5–21.

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121–1133.

Rus, V., Niraula, N., & Banjade, R. (2013). Similarity measures based on latent dirichlet allocation. In *International conference on intelligent text processing and computational linguistics* (pp. 459–470).

Salazar, A. J. (1996). An analysis of the development and evolution of roles in the small group. *Small Group Research*, *27*(4), 475–503.

Scissors, L. E., Gill, A. J., Geraghty, K., & Gergle, D. (2009). In cmc we trust: The role of similarity. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 527–536).

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society*, *5*(1), 1–23.

Setiawan, R., Budiharto, W., Kartowisastro, I. H., & Prabowo, H. (2020). Finding model through latent semantic approach to reveal the topic of discussion in discussion

forum. *Education and Information Technologies*, *25*(1), 31–50.

Shah, D. (2021). By the numbers: Moocs in 2020. *Class Central Report, found online at https://www.classcentral.com/report/mooc-stats-2021*.

Shahhoseiny, H. (2013). Differences between language and linguistic in the elt classroom. *Theory & Practice in Language Studies*, *3*(12).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shao, Z., Yang, S., Gao, F., Zhou, K., & Lin, P. (2017). A new electricity price prediction strategy using mutual information-based svm-rfe classification. *Renewable and Sustainable Energy Reviews*, *70*, 330–341.

Sharkey, M., & Sanders, R. (2014). A process for predicting mooc attrition. In *Proceedings of the emnlp 2014 workshop on analysis of large scale social interaction in moocs* (pp. 50–54).

Sherblom, J. C. (1990). Organization involvement expressed through pronoun use in computer mediated communication. *Communication Research Reports*, *7*(1), 45–50.

Shipp, A. J., Edwards, J. R., & Lambert, L. S. (2009). Conceptualization and measurement of temporal focus: The subjective experience of the past, present, and future. *Organizational behavior and human decision processes*, *110*(1), 1–22.

Sinclair, J., & Cardew-Hall, M. (2008). The folksonomy tag cloud: when is it useful? *Journal of Information Science*, *34*(1), 15–29.

Smith-Keiling, B. L., & Hyun, H. I. F. (2019). Applying a computer-assisted tool for semantic analysis of writing: uses for stem and ell. *Journal of microbiology & biology education*, *20*(1), 70.

Smith-Keiling, B. L., Swanson, L. K., & Dehnbostel, J. M. (2018). Interventions for supporting and assessing science writing communication: cases of asian english language learners. *Journal of microbiology & biology education*, *19*(1), 19–1.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth acm conference on conference on information and knowledge management* (pp. 623–632).

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, *25*.

Soller, A. (2001). Supporting social interaction in an intelligent collaborative learning system. *International journal of artificial intelligence in education*, *12*(1), 40–62.

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

Sridhar, D., Getoor, L., & Walker, M. (2014). Collective stance classification of posts in online debate forums. In *Proceedings of the joint workshop on social dynamics and personal attributes in social media* (pp. 109–117).

Srinivas, H. (2011). What is collaborative learning. *The Global Development Research Center, Kobe*.

Stahl, G. (2004). Building collaborative knowing. In *What we know about cscl* (pp. 53–85). Springer.

Stasko, J., Görg, C., & Liu, Z. (2008). Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, *7*(2), 118–132.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 952–961).

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, *63*(4), 517–522.

Stone, L. D., & Pennebaker, J. W. (2002). Trauma in real time: Talking and avoiding online conversations about the death of princess diana. *Basic and applied social psychology*, *24*(3), 173–183.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research.* Sage publications.

Strijbos, J.-W., & De Laat, M. F. (2010). Developing the role concept for computer-supported collaborative learning: An explorative synthesis. *Computers in human behavior*, *26*(4), 495–505.

Strijbos, J.-W., & Weinberger, A. (2010). Emerging and scripted roles in computer-supported collaborative learning. *Computers in Human Behavior*, *26*(4), 491–494.

Tan, C.-M., Wang, Y.-F., & Lee, C.-D. (2002). The use of bigrams to enhance text categorization. *Information processing & management*, *38*(4), 529–546.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social*

psychology, *29*(1), 24–54.

Teppo, A. R. (2015). Grounded theory methods. In *Approaches to qualitative research in mathematics education* (pp. 3–21). Springer.

Tetlock, P. E. (1981). Pre-to postelection shifts in presidential rhetoric: Impression management or cognitive adjustment. *Journal of Personality and Social Psychology*, *41*(2), 207.

Tuan, A., Aleti, T., Pallant, J., & van Laer, T. (2019). Tweeting with the stars: Analyzing linguistic styles of celebrities' tweet and their effect on consumer word of mouth. *ACR North American Advances*.

Turney, P., Neuman, Y., Assaf, D., & Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 680–690).

Van Dijk, T. A. (1997). *Discourse as structure and process* (Vol. 1). Sage.

Vardi, M. (2012). *Will moocs destroy academia? communica tions of the acm, 55 (11).* nov.

Veldhuis-Diermanse, A. E. (2002). Csclearning?: participation, learning, activities and knowledge construction in computer-supported collaborative learning in higher education.

Volet, S., Vauras, M., Salo, A.-E., & Khosa, D. (2017). Individual contributions in student-led collaborative learning: Insights from two analytical approaches to explain the quality of group outcome. *Learning and Individual Differences*, *53*, 79–92.

Vollstedt, M., & Rezat, S. (2019). An introduction to grounded theory with a special focus on axial coding and the coding paradigm. *Compendium for early career researchers in mathematics education*, *13*, 81–100.

Vytasek, J. M., Wise, A. F., & Woloshen, S. (2017). Topic models to support instructors in mooc forums. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 610–611).

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating how student's cognitive behavior in mooc discussion forums affect learning gains. *International Educational Data Mining Society*.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *American Educational Research Journal*, *35*(4), 607–651.

Webb, N. M., Troper, J. D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of educational psychology*, *87*(3), 406.

Weigand, H., Goldkuhl, G., & de Moor, A. (2003). *Proceedings of the 8th international working conference on the language-action perspective on communication modeling (lap 2003)* (Tech. Rep.). Tilburg University, School of Economics and Management.

Weinberger, A., Stegmann, K., & Fischer, F. (2010). Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). *Computers in Human behavior*, *26*(4), 506–515.

Weintraub, W. (1989). *Verbal behavior in everyday life.* Springer Publishing Co.

Wen, M., Yang, D., & Rosé, C. (2014). Linguistic reflections of student engagement in massive open online courses. In *Proceedings of the international aaai conference on web and social media* (Vol. 8).

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

Westerlund, M., Mahmood, Z., Leminen, S., & Rajahonka, M. (2019). Topic modelling analysis of online reviews: Indian restaurants at amazon. com. In *Ispim conference proceedings* (pp. 1–14).

Winograd, T. (1987). A language/action perspective on the design of cooperative work. *Human–Computer Interaction*, *3*, 3-30.

Winship, C., & Mandel, M. (1983). Roles and positions: A critique and extension of the blockmodeling approach. *Sociological methodology*, *14*, 314–344.

Wise, A. F., Cui, Y., Jin, W., & Vytasek, J. (2017). Mining for gold: Identifying content-related mooc discussion threads across domains through linguistic modeling. *The Internet and Higher Education*, *32*, 11–28.

Wise, A. F., Cui, Y., & Vytasek, J. (2016). Bringing order to chaos in mooc discussion forums with content-related thread identification. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 188–197).

Wong, A. W., Wong, K., & Hindle, A. (2019). Tracing forum posts to mooc content using topic analysis. *arXiv preprint arXiv:1904.07307*.

Wong, J.-S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015). An analysis of mooc discussion forum interactions from the most active users. In *International conference on social computing, behavioral-cultural modeling, and prediction* (pp. 452–457).

Wu, H., Zhang, X., Xie, H., Kuang, Y., & Ouyang, G. (2013). Classification of solder joint using feature selection based on bayes and support vector machine. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, *3*(3), 516–522.

Wu, Y., Wei, F., Liu, S., Au, N., Cui, W., Zhou, H., & Qu, H. (2010). Opinionseer: interactive visualization of hotel customer feedback. *IEEE transactions on visualization and computer graphics*, *16*(6), 1109–1118.

Wu, Y., & Zhang, A. (2004). Feature selection for classifying high-dimensional numerical data. In *Proceedings of the 2004 ieee computer society conference on computer vision and pattern recognition, 2004. cvpr 2004.* (Vol. 2, pp. II–II).

Wulf, J., Blohm, I., Leimeister, J. M., & Brenner, W. (2014). Massive open online courses. *Business & Information Systems Engineering*, *6*(2), 111–114.

Xie, W., Zhu, F., Jiang, J., Lim, E.-P., & Wang, K. (2016). Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 2216–2229.

Xu, B., & Yang, D. (2015). Study partners recommendation for xmoocs learners. *Computational intelligence and neuroscience*, *2015*.

Yan, W., Dowell, N., Holman, C., Welsh, S. S., Choi, H., & Brooks, C. (2019). Exploring learner engagement patterns in teach-outs using topic, sentiment and on-topicness to reflect on pedagogy. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 180–184).

Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 nips data-driven education workshop* (Vol. 11, p. 14).

Yang, D., Wen, M., Howley, I., Kraut, R., & Rose, C. (2015). Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the second (2015) acm conference on learning@ scale* (pp. 121–130).

Yang, D., Wen, M., & Rose, C. (2014). Peer influence on attrition in massively open online courses. In *Educational data mining 2014*.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141–2155.

Youssef, M., Mohammed, S., Hamada, E. K., & Wafaa, B. F. (2019). A predictive approach based on efficient feature selection and learning algorithms' competition: Case of learners' dropout in moocs. *Education and Information Technologies*, *24*(6), 3591–3618.

Zarra, T., Chiheb, R., Faizi, R., & El Afia, A. (2018). Student interactions in online discussion forums: Visual analysis with lda topic models. In *Proceedings of the international conference on learning and optimization algorithms: Theory and applications* (pp. 1–5).

Zeng, X., & Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(1), 1–12.

Zeng, Z., Chaturvedi, S., & Bhat, S. (2017). Learner affect through the looking glass: Characterization and detection of confusion in online courses. *International Educational Data Mining Society*.

Zhang, J. (1998). A distributed representation approach to group problem solving. *Journal of the American Society for Information Science*, *49*(9), 801–809.

Zheng, S., Rosson, M. B., Shih, P. C., & Carroll, J. M. (2015). Understanding student motivation, behaviors and perceptions in moocs. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing* (pp. 1882–1895).

Zhu, M., Herring, S. C., & Bonk, C. J. (2019). Exploring presence in online learning through three forms of computer-mediated discourse analysis. *Distance Education*, *40*(2), 205–225.

Zong, S., Ritter, A., & Hovy, E. (2020). Measuring forecasting skill from text. *arXiv preprint arXiv:2006.07425*.

Appendix A

Mark Scheme

M10 – Mark Scheme 10

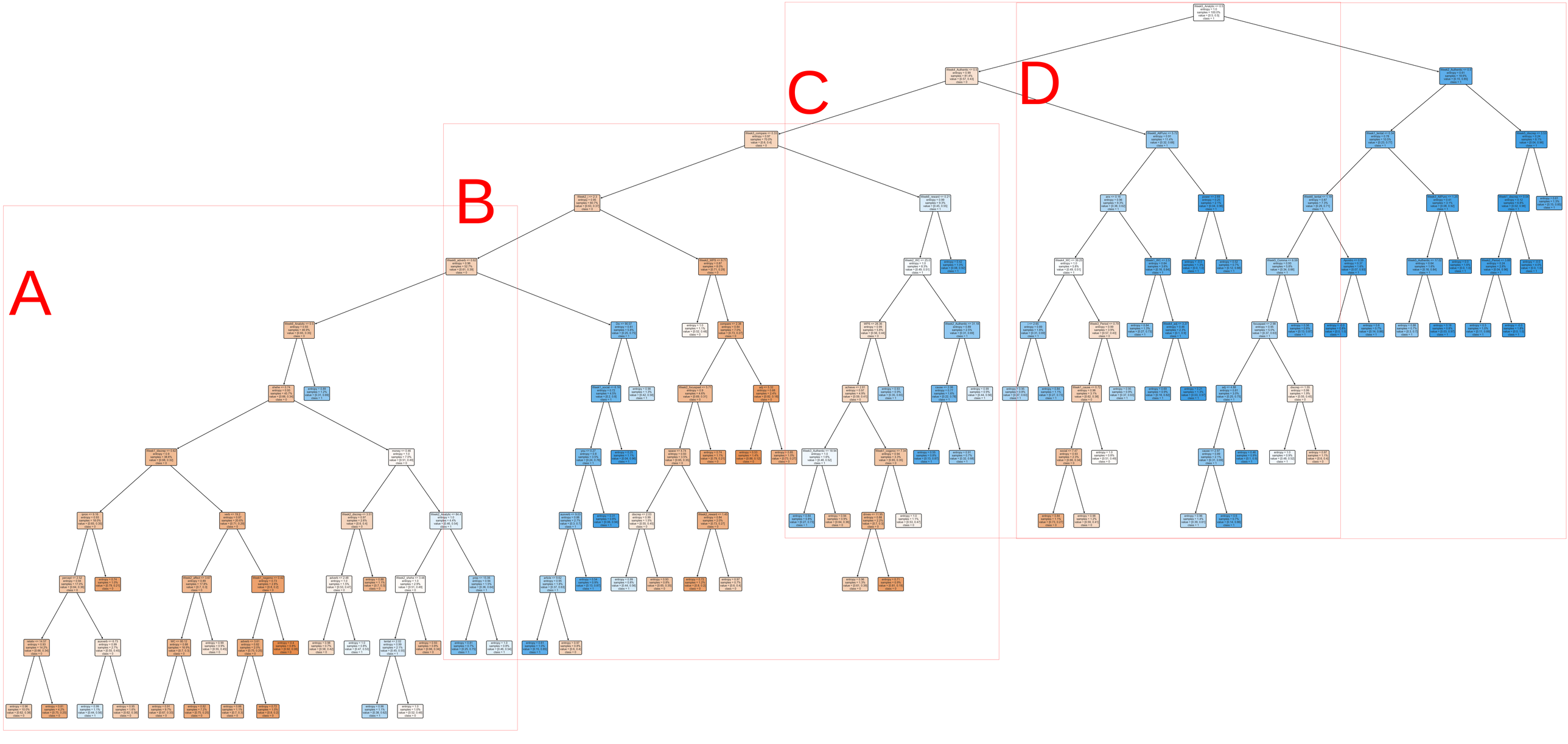
This scheme is the norm for courses in undergraduate and postgraduate coursework programs.
This scheme applies to the **course final result**.

Grade	Grade reflects following criteria for allocation of grade:	Reported outcome
Fail No Submission	No work submitted for assessment	FNS
Fail	A mark between 1-49	F
Pass	A mark between 50-64	50-64 P
Credit	A mark between 65-74	65-74 C
Distinction	A mark between 75-84	75-84 D
High Distinction	A mark between 85-100	85-100 HD

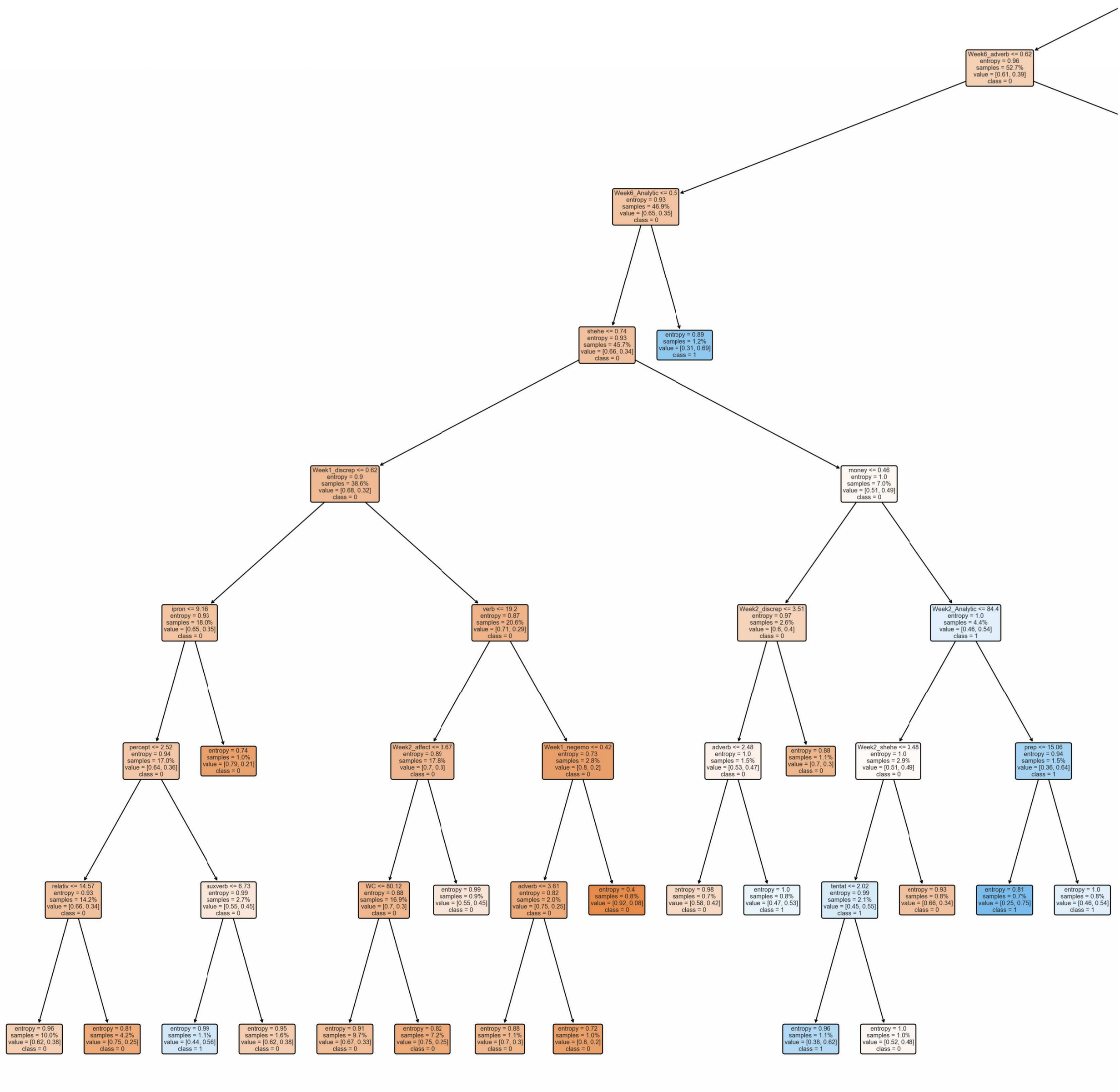
FIGURE 1: Mark Scheme

Appendix B

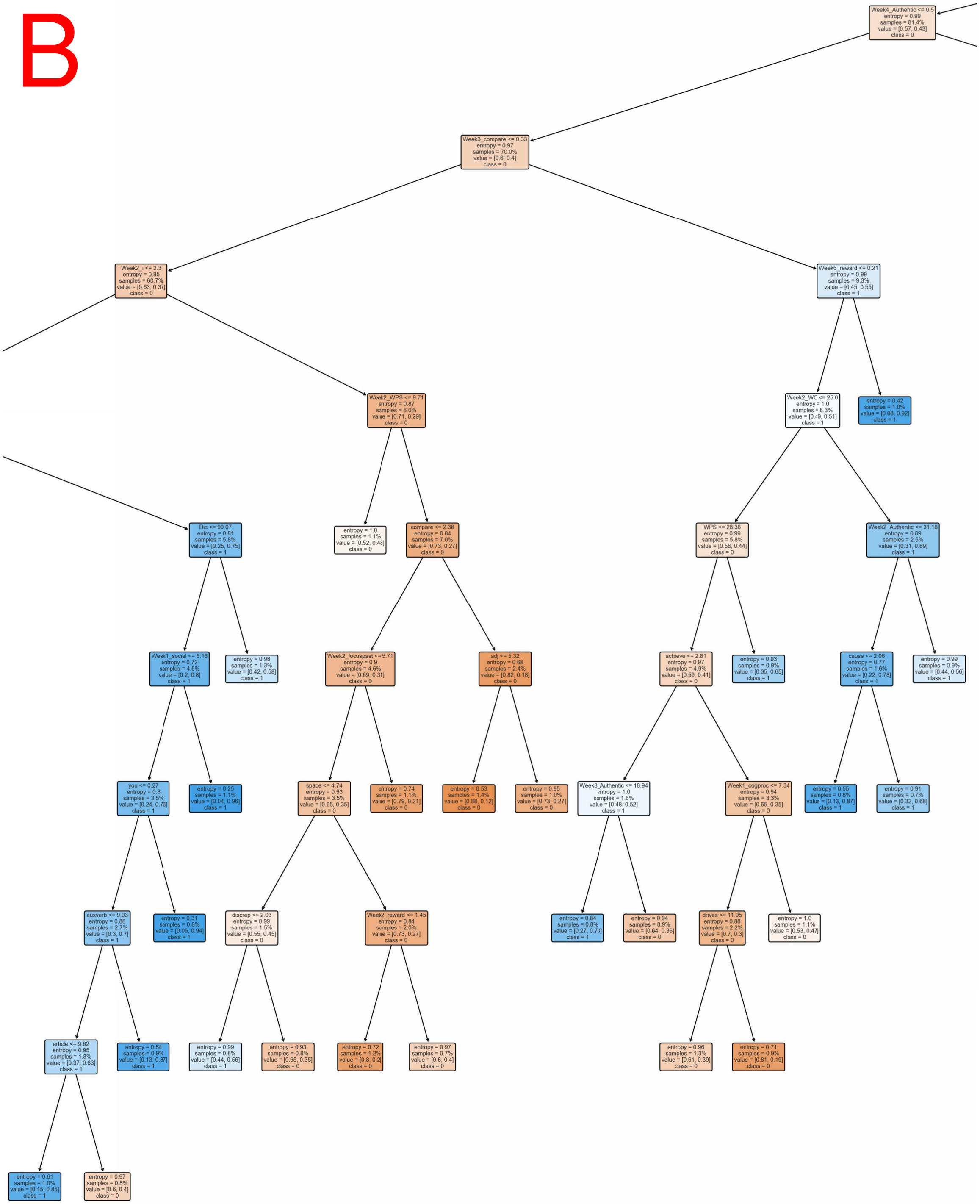
Decision Tree


A
B
C
D

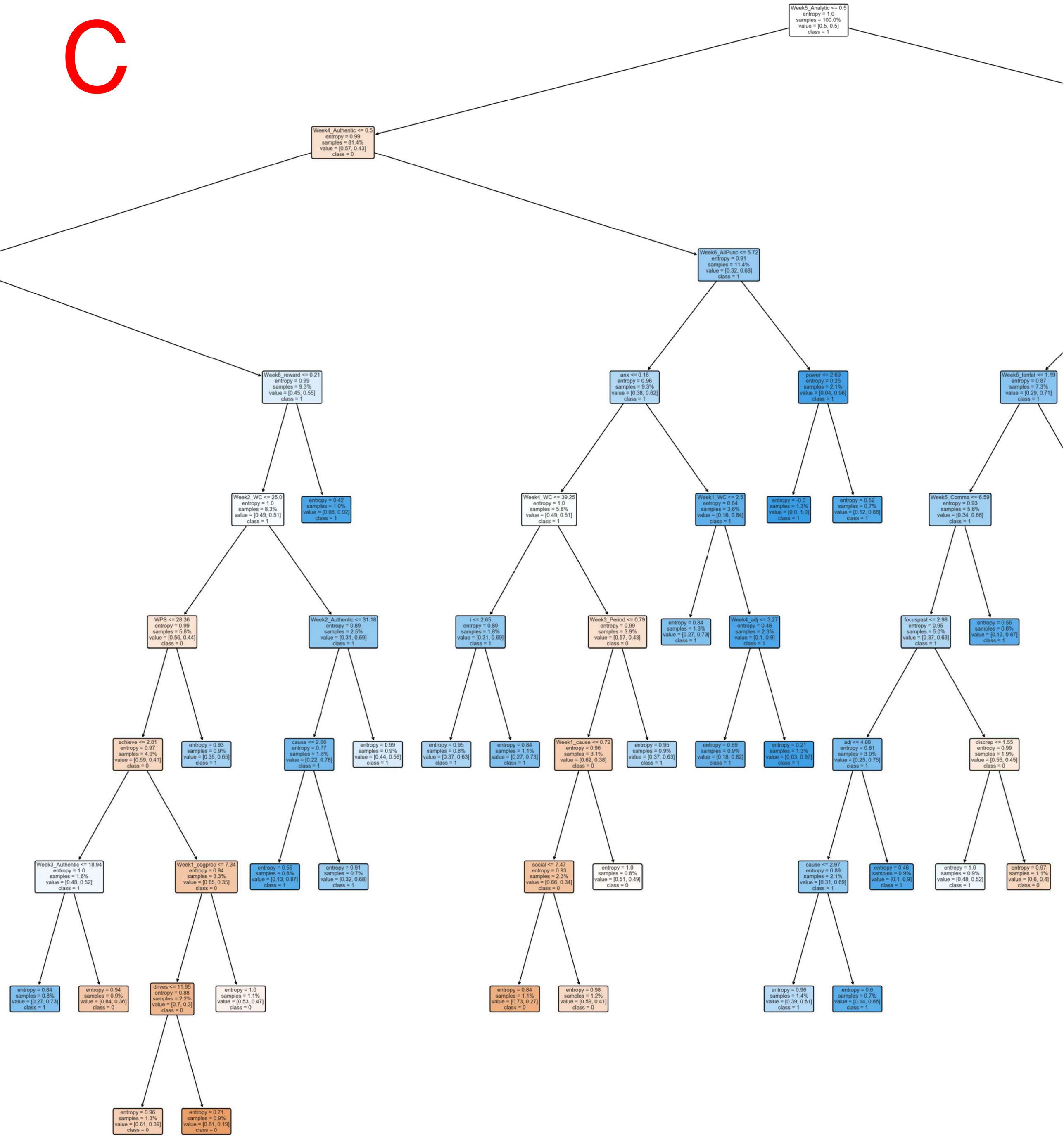
A


Week6_adverb <= 0.62
entropy = 0.96
samples = 52.7%
value = [0.61, 0.39]
class = 0
Week6_Analytic <= 0.5
entropy = 0.93
samples = 46.9%
value = [0.65, 0.35]
class = 0
entropy = 0.89
samples = 1.2%
value = [0.31, 0.69]
class = 1
shehe <= 0.74
entropy = 0.93
samples = 45.7%
value = [0.66, 0.34]
class = 0
Week1_discrep <= 0.62
entropy = 0.9
samples = 38.6%
value = [0.68, 0.32]
class = 0
money <= 0.46
entropy = 1.0
samples = 7.0%
value = [0.51, 0.49]
class = 0
ipron <= 9.16
entropy = 0.93
samples = 18.0%
value = [0.65, 0.35]
class = 0
verb <= 19.2
entropy = 0.87
samples = 20.6%
value = [0.71, 0.29]
class = 0
Week2_discrep <= 3.51
entropy = 0.97
samples = 2.6%
value = [0.6, 0.4]
class = 0
Week2_Analytic <= 84.4
entropy = 1.0
samples = 4.4%
value = [0.46, 0.54]
class = 1
percept <= 2.52
entropy = 0.94
samples = 17.0%
value = [0.64, 0.36]
class = 0
entropy = 0.74
samples = 1.0%
value = [0.79, 0.21]
class = 0
Week2_affect <= 3.67
entropy = 0.89
samples = 17.8%
value = [0.7, 0.3]
class = 0
Week1_negemo <= 0.42
entropy = 0.73
samples = 2.8%
value = [0.8, 0.2]
class = 0
adverb <= 2.48
entropy = 1.0
samples = 1.5%
value = [0.53, 0.47]
class = 0
entropy = 0.88
samples = 1.1%
value = [0.7, 0.3]
class = 0
Week2_shehe <= 3.48
entropy = 1.0
samples = 2.9%
value = [0.51, 0.49]
class = 0
prep <= 15.06
entropy = 0.94
samples = 1.5%
value = [0.36, 0.64]
class = 1
relativ <= 14.57
entropy = 0.93
samples = 14.2%
value = [0.66, 0.34]
class = 0
auxverb <= 6.73
entropy = 0.99
samples = 2.7%
value = [0.55, 0.45]
class = 0
WC <= 80.12
entropy = 0.88
samples = 16.9%
value = [0.7, 0.3]
class = 0
entropy = 0.99
samples = 0.9%
value = [0.55, 0.45]
class = 0
adverb <= 3.61
entropy = 0.82
samples = 2.0%
value = [0.75, 0.25]
class = 0
entropy = 0.4
samples = 0.8%
value = [0.92, 0.08]
class = 0
entropy = 0.98
samples = 0.7%
value = [0.58, 0.42]
class = 0
entropy = 1.0
samples = 0.8%
value = [0.47, 0.53]
class = 1
tentat <= 2.02
entropy = 0.99
samples = 2.1%
value = [0.45, 0.55]
class = 1
entropy = 0.93
samples = 0.8%
value = [0.66, 0.34]
class = 0
entropy = 0.81
samples = 0.7%
value = [0.25, 0.75]
class = 1
entropy = 1.0
samples = 0.8%
value = [0.46, 0.54]
class = 1
entropy = 0.96
samples = 10.0%
value = [0.62, 0.38]
class = 0
entropy = 0.81
samples = 4.2%
value = [0.75, 0.25]
class = 0
entropy = 0.99
samples = 1.1%
value = [0.44, 0.56]
class = 1
entropy = 0.95
samples = 1.6%
value = [0.62, 0.38]
class = 0
entropy = 0.91
samples = 9.7%
value = [0.67, 0.33]
class = 0
entropy = 0.82
samples = 7.2%
value = [0.75, 0.25]
class = 0
entropy = 0.88
samples = 1.1%
value = [0.7, 0.3]
class = 0
entropy = 0.72
samples = 1.0%
value = [0.8, 0.2]
class = 0
entropy = 0.96
samples = 1.1%
value = [0.38, 0.62]
class = 1
entropy = 1.0
samples = 1.0%
value = [0.52, 0.48]
class = 0

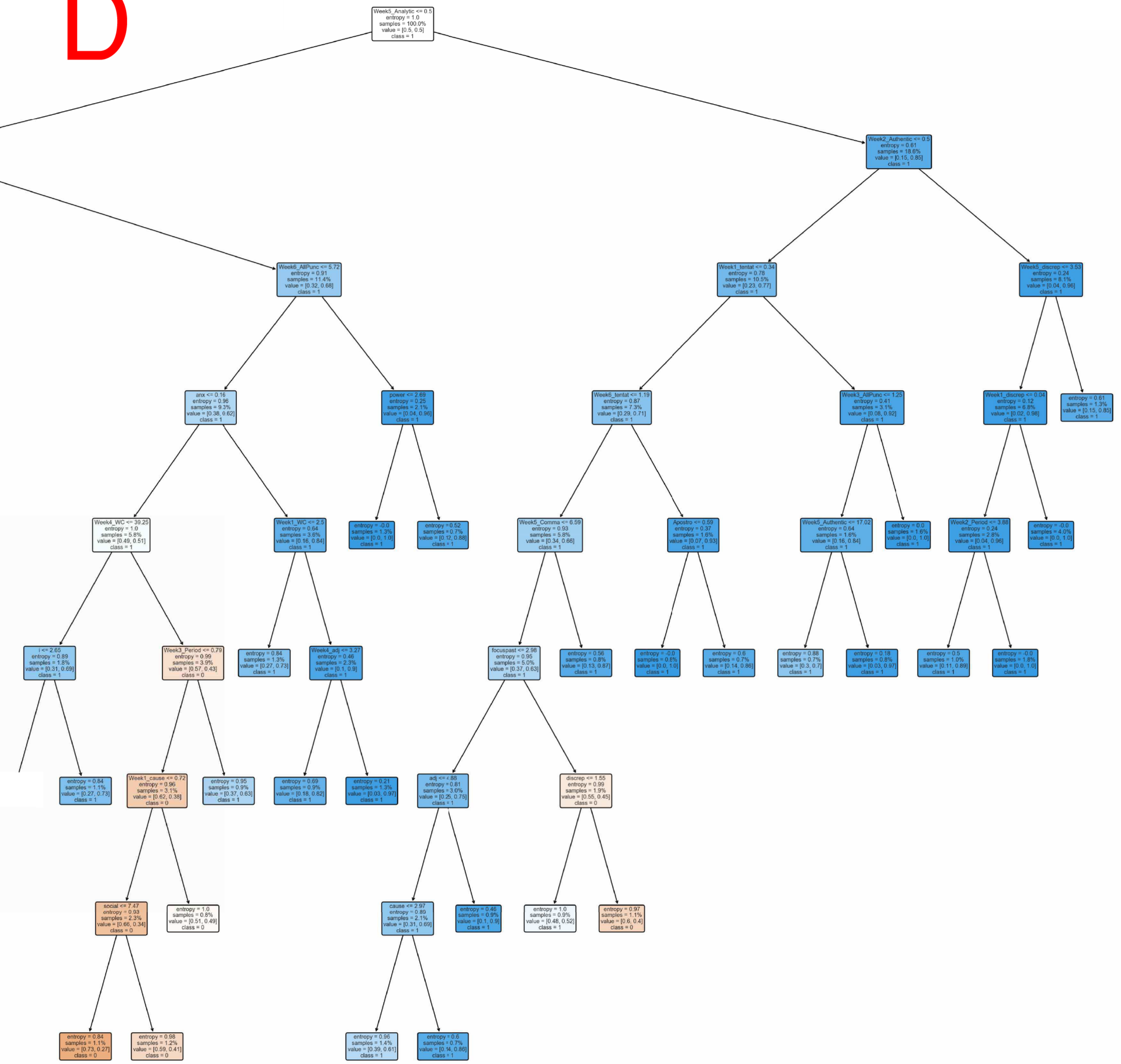
B


Week4_Authentic <= 0.5
entropy = 0.99
samples = 81.4%
value = [0.57, 0.43]
class = 0
Week3_compare <= 0.33
entropy = 0.97
samples = 70.0%
value = [0.6, 0.4]
class = 0
Week2_i <= 2.3
entropy = 0.95
samples = 60.7%
value = [0.63, 0.37]
class = 0
Week6_reward <= 0.21
entropy = 0.99
samples = 9.3%
value = [0.45, 0.55]
class = 1
Week2_WPS <= 9.71
entropy = 0.87
samples = 8.0%
value = [0.71, 0.29]
class = 0
Week2_WC <= 25.0
entropy = 1.0
samples = 8.3%
value = [0.49, 0.51]
class = 1
entropy = 0.42
samples = 1.0%
value = [0.08, 0.92]
class = 1
Dic <= 90.07
entropy = 0.81
samples = 5.8%
value = [0.25, 0.75]
class = 1
entropy = 1.0
samples = 1.1%
value = [0.52, 0.43]
class = 0
compare <= 2.38
entropy = 0.84
samples = 7.0%
value = [0.73, 0.27]
class = 0
WPS <= 28.36
entropy = 0.99
samples = 5.8%
value = [0.56, 0.44]
class = 0
Week2_Authentic <= 31.18
entropy = 0.89
samples = 2.5%
value = [0.31, 0.69]
class = 1
Week1_social <= 6.16
entropy = 0.72
samples = 4.5%
value = [0.2, 0.8]
class = 1
entropy = 0.98
samples = 1.3%
value = [0.42, 0.58]
class = 1
Week2_focuspast <= 5.71
entropy = 0.9
samples = 4.6%
value = [0.69, 0.31]
class = 0
adj <= 5.32
entropy = 0.68
samples = 2.4%
value = [0.82, 0.18]
class = 0
achieve <= 2.81
entropy = 0.97
samples = 4.9%
value = [0.59, 0.41]
class = 0
entropy = 0.93
samples = 0.9%
value = [0.35, 0.65]
class = 1
cause <= 2.06
entropy = 0.77
samples = 1.6%
value = [0.22, 0.78]
class = 1
entropy = 0.99
samples = 0.9%
value = [0.44, 0.56]
class = 1
you <= 0.27
entropy = 0.8
samples = 3.5%
value = [0.24, 0.76]
class = 1
entropy = 0.25
samples = 1.1%
value = [0.04, 0.96]
class = 1
space <= 4.74
entropy = 0.93
samples = 3.5%
value = [0.65, 0.35]
class = 0
entropy = 0.74
samples = 1.1%
value = [0.79, 0.21]
class = 0
entropy = 0.53
samples = 1.4%
value = [0.88, 0.12]
class = 0
entropy = 0.85
samples = 1.0%
value = [0.73, 0.27]
class = 0
Week3_Authentic <= 18.94
entropy = 1.0
samples = 1.6%
value = [0.48, 0.52]
class = 1
Week1_cogproc <= 7.34
entropy = 0.94
samples = 3.3%
value = [0.65, 0.35]
class = 0
entropy = 0.55
samples = 0.8%
value = [0.13, 0.87]
class = 1
entropy = 0.91
samples = 0.7%
value = [0.32, 0.68]
class = 1
auxverb <= 9.03
entropy = 0.88
samples = 2.7%
value = [0.3, 0.7]
class = 1
entropy = 0.31
samples = 0.8%
value = [0.06, 0.94]
class = 1
discrep <= 2.03
entropy = 0.99
samples = 1.5%
value = [0.55, 0.45]
class = 0
Week2_reward <= 1.45
entropy = 0.84
samples = 2.0%
value = [0.73, 0.27]
class = 0
entropy = 0.84
samples = 0.8%
value = [0.27, 0.73]
class = 1
entropy = 0.94
samples = 0.9%
value = [0.64, 0.36]
class = 0
drives <= 11.95
entropy = 0.88
samples = 2.2%
value = [0.7, 0.3]
class = 0
entropy = 1.0
samples = 1.1%
value = [0.53, 0.47]
class = 0
article <= 9.62
entropy = 0.95
samples = 1.8%
value = [0.37, 0.63]
class = 1
entropy = 0.54
samples = 0.9%
value = [0.13, 0.87]
class = 1
entropy = 0.99
samples = 0.8%
value = [0.44, 0.56]
class = 1
entropy = 0.93
samples = 0.8%
value = [0.65, 0.35]
class = 0
entropy = 0.72
samples = 1.2%
value = [0.8, 0.2]
class = 0
entropy = 0.97
samples = 0.7%
value = [0.6, 0.4]
class = 0
entropy = 0.96
samples = 1.3%
value = [0.61, 0.39]
class = 0
entropy = 0.71
samples = 0.9%
value = [0.81, 0.19]
class = 0
entropy = 0.61
samples = 1.0%
value = [0.15, 0.85]
class = 1
entropy = 0.97
samples = 0.8%
value = [0.6, 0.4]
class = 0

C


Week5_Analytic <= 0.5
entropy = 1.0
samples = 100.0%
value = [0.5, 0.5]
class = 1
Week4_Authentic <= 0.5
entropy = 0.99
samples = 81.4%
value = [0.57, 0.43]
class = 0
Week6_AllPunc <= 5.72
entropy = 0.91
samples = 11.4%
value = [0.32, 0.68]
class = 1
Week6_reward <= 0.21
entropy = 0.99
samples = 9.3%
value = [0.45, 0.55]
class = 1
anx <= 0.16
entropy = 0.96
samples = 9.3%
value = [0.38, 0.62]
class = 1
power <= 2.69
entropy = 0.25
samples = 2.1%
value = [0.04, 0.96]
class = 1
Week6_tentat <= 1.19
entropy = 0.87
samples = 7.3%
value = [0.29, 0.71]
class = 1
Week2_WC <= 25.0
entropy = 1.0
samples = 8.3%
value = [0.49, 0.51]
class = 1
entropy = 0.42
samples = 1.0%
value = [0.08, 0.92]
class = 1
Week4_WC <= 39.25
entropy = 1.0
samples = 5.8%
value = [0.49, 0.51]
class = 1
Week1_WC <= 2.5
entropy = 0.64
samples = 3.6%
value = [0.16, 0.84]
class = 1
entropy = -0.0
samples = 1.3%
value = [0.0, 1.0]
class = 1
entropy = 0.52
samples = 0.7%
value = [0.12, 0.88]
class = 1
Week5_Comma <= 6.59
entropy = 0.93
samples = 5.8%
value = [0.34, 0.66]
class = 1
WPS <= 28.36
entropy = 0.99
samples = 5.8%
value = [0.56, 0.44]
class = 0
Week2_Authentic <= 31.18
entropy = 0.89
samples = 2.5%
value = [0.31, 0.69]
class = 1
i <= 2.65
entropy = 0.89
samples = 1.8%
value = [0.31, 0.69]
class = 1
Week3_Period <= 0.79
entropy = 0.99
samples = 3.9%
value = [0.57, 0.43]
class = 0
entropy = 0.84
samples = 1.3%
value = [0.27, 0.73]
class = 1
Week4_adj <= 3.27
entropy = 0.46
samples = 2.3%
value = [0.1, 0.9]
class = 1
focuspast <= 2.98
entropy = 0.95
samples = 5.0%
value = [0.37, 0.63]
class = 1
entropy = 0.56
samples = 0.8%
value = [0.13, 0.87]
class = 1
achieve <= 2.81
entropy = 0.97
samples = 4.9%
value = [0.59, 0.41]
class = 0
entropy = 0.93
samples = 0.9%
value = [0.35, 0.65]
class = 1
cause <= 2.06
entropy = 0.77
samples = 1.6%
value = [0.22, 0.78]
class = 1
entropy = 0.99
samples = 0.9%
value = [0.44, 0.56]
class = 1
entropy = 0.95
samples = 0.8%
value = [0.37, 0.63]
class = 1
entropy = 0.84
samples = 1.1%
value = [0.27, 0.73]
class = 1
Week1_cause <= 0.72
entropy = 0.96
samples = 3.1%
value = [0.62, 0.38]
class = 0
entropy = 0.95
samples = 0.9%
value = [0.37, 0.63]
class = 1
entropy = 0.69
samples = 0.9%
value = [0.18, 0.82]
class = 1
entropy = 0.21
samples = 1.3%
value = [0.03, 0.97]
class = 1
adj <= 4.88
entropy = 0.81
samples = 3.0%
value = [0.25, 0.75]
class = 1
discrep <= 1.55
entropy = 0.99
samples = 1.9%
value = [0.55, 0.45]
class = 0
Week3_Authentic <= 18.94
entropy = 1.0
samples = 1.6%
value = [0.48, 0.52]
class = 1
Week1_cogproc <= 7.34
entropy = 0.94
samples = 3.3%
value = [0.65, 0.35]
class = 0
entropy = 0.55
samples = 0.8%
value = [0.13, 0.87]
class = 1
entropy = 0.91
samples = 0.7%
value = [0.32, 0.68]
class = 1
social <= 7.47
entropy = 0.93
samples = 2.3%
value = [0.66, 0.34]
class = 0
entropy = 1.0
samples = 0.8%
value = [0.51, 0.49]
class = 0
cause <= 2.97
entropy = 0.89
samples = 2.1%
value = [0.31, 0.69]
class = 1
entropy = 0.46
samples = 0.9%
value = [0.1, 0.9]
class = 1
entropy = 1.0
samples = 0.9%
value = [0.48, 0.52]
class = 1
entropy = 0.97
samples = 1.1%
value = [0.6, 0.4]
class = 0
entropy = 0.84
samples = 0.8%
value = [0.27, 0.73]
class = 1
entropy = 0.94
samples = 0.9%
value = [0.64, 0.36]
class = 0
drives <= 11.95
entropy = 0.88
samples = 2.2%
value = [0.7, 0.3]
class = 0
entropy = 1.0
samples = 1.1%
value = [0.53, 0.47]
class = 0
entropy = 0.84
samples = 1.1%
value = [0.73, 0.27]
class = 0
entropy = 0.98
samples = 1.2%
value = [0.59, 0.41]
class = 0
entropy = 0.96
samples = 1.4%
value = [0.39, 0.61]
class = 1
entropy = 0.6
samples = 0.7%
value = [0.14, 0.86]
class = 1
entropy = 0.96
samples = 1.3%
value = [0.61, 0.39]
class = 0
entropy = 0.71
samples = 0.9%
value = [0.81, 0.19]
class = 0

D


Week5_Analytic <= 0.5
entropy = 1.0
samples = 100.0%
value = [0.5, 0.5]
class = 1
Week6_AllPunc <= 5.72
entropy = 0.91
samples = 11.4%
value = [0.32, 0.68]
class = 1
anx <= 0.16
entropy = 0.96
samples = 9.3%
value = [0.38, 0.62]
class = 1
power <= 2.69
entropy = 0.25
samples = 2.1%
value = [0.04, 0.96]
class = 1
Week4_WC <= 39.25
entropy = 1.0
samples = 5.8%
value = [0.49, 0.51]
class = 1
Week1_WC <= 2.5
entropy = 0.64
samples = 3.6%
value = [0.16, 0.84]
class = 1
entropy = -0.0
samples = 1.3%
value = [0.0, 1.0]
class = 1
entropy = 0.52
samples = 0.7%
value = [0.12, 0.88]
class = 1
i <= 2.65
entropy = 0.89
samples = 1.8%
value = [0.31, 0.69]
class = 1
Week3_Period <= 0.79
entropy = 0.99
samples = 3.9%
value = [0.57, 0.43]
class = 0
entropy = 0.84
samples = 1.3%
value = [0.27, 0.73]
class = 1
Week4_adj <= 3.27
entropy = 0.46
samples = 2.3%
value = [0.1, 0.9]
class = 1
entropy = 0.84
samples = 1.1%
value = [0.27, 0.73]
class = 1
Week1_cause <= 0.72
entropy = 0.96
samples = 3.1%
value = [0.62, 0.38]
class = 0
entropy = 0.95
samples = 0.9%
value = [0.37, 0.63]
class = 1
entropy = 0.69
samples = 0.9%
value = [0.18, 0.82]
class = 1
entropy = 0.21
samples = 1.3%
value = [0.03, 0.97]
class = 1
social <= 7.47
entropy = 0.93
samples = 2.3%
value = [0.66, 0.34]
class = 0
entropy = 1.0
samples = 0.8%
value = [0.51, 0.49]
class = 0
entropy = 0.84
samples = 1.1%
value = [0.73, 0.27]
class = 0
entropy = 0.98
samples = 1.2%
value = [0.59, 0.41]
class = 0
Week2_Authentic <= 0.5
entropy = 0.61
samples = 18.6%
value = [0.15, 0.85]
class = 1
Week1_tentat <= 0.34
entropy = 0.78
samples = 10.5%
value = [0.23, 0.77]
class = 1
Week5_discrep <= 3.53
entropy = 0.24
samples = 8.1%
value = [0.04, 0.96]
class = 1
Week6_tentat <= 1.19
entropy = 0.87
samples = 7.3%
value = [0.29, 0.71]
class = 1
Week3_AllPunc <= 1.25
entropy = 0.41
samples = 3.1%
value = [0.08, 0.92]
class = 1
Week1_discrep <= 0.04
entropy = 0.12
samples = 6.8%
value = [0.02, 0.98]
class = 1
entropy = 0.61
samples = 1.3%
value = [0.15, 0.85]
class = 1
Week5_Comma <= 6.59
entropy = 0.93
samples = 5.8%
value = [0.34, 0.66]
class = 1
Apostro <= 0.59
entropy = 0.37
samples = 1.6%
value = [0.07, 0.93]
class = 1
Week5_Authentic <= 17.02
entropy = 0.64
samples = 1.6%
value = [0.16, 0.84]
class = 1
entropy = 0.0
samples = 1.6%
value = [0.0, 1.0]
class = 1
Week2_Period <= 3.88
entropy = 0.24
samples = 2.8%
value = [0.04, 0.96]
class = 1
entropy = -0.0
samples = 4.0%
value = [0.0, 1.0]
class = 1
focuspast <= 2.98
entropy = 0.95
samples = 5.0%
value = [0.37, 0.63]
class = 1
entropy = 0.56
samples = 0.8%
value = [0.13, 0.87]
class = 1
entropy = -0.0
samples = 0.6%
value = [0.0, 1.0]
class = 1
entropy = 0.6
samples = 0.7%
value = [0.14, 0.86]
class = 1
entropy = 0.88
samples = 0.7%
value = [0.3, 0.7]
class = 1
entropy = 0.18
samples = 0.8%
value = [0.03, 0.97]
class = 1
entropy = 0.5
samples = 1.0%
value = [0.11, 0.89]
class = 1
entropy = -0.0
samples = 1.8%
value = [0.0, 1.0]
class = 1
adj <= 4.88
entropy = 0.81
samples = 3.0%
value = [0.25, 0.75]
class = 1
discrep <= 1.55
entropy = 0.99
samples = 1.9%
value = [0.55, 0.45]
class = 0
cause <= 2.97
entropy = 0.89
samples = 2.1%
value = [0.31, 0.69]
class = 1
entropy = 0.46
samples = 0.9%
value = [0.1, 0.9]
class = 1
entropy = 1.0
samples = 0.9%
value = [0.48, 0.52]
class = 1
entropy = 0.97
samples = 1.1%
value = [0.6, 0.4]
class = 0
entropy = 0.96
samples = 1.4%
value = [0.39, 0.61]
class = 1
entropy = 0.6
samples = 0.7%
value = [0.14, 0.86]
class = 1