Predicting Paid Certification in Massive Open Online Courses

MOHAMMAD ABDULLAH ALSHEHRI

Predicting Paid Certification in Massive Open Online Courses

Mohammad Alshehri

A Thesis Presented for the Degree of Doctor of Philosophy in Computer Science

Supervised by:

Prof. Alexandra I. Cristea

Artificial Intelligence and Human Systems Research Group

University of Durham

United Kingdom

2023

Dedication

To my father, who has been ICU hospitalised many times while writing this thesis. Never give up, Dad … God gives his hardest battles to his toughest soldiers.

Predicting Paid Certification in Massive Open Online Courses

Mohammad Alshehri

Submitted for the Degree of Doctor of Philosophy

2023

Abstract

Massive open online courses (MOOCs) have been proliferating because of the free or low-cost offering of content for learners, attracting the attention of many stakeholders across the entire educational landscape. Since 2012, coined as "the Year of the MOOCs", several platforms have gathered millions of learners in just a decade. Nevertheless, the certification rate of both free and paid courses has been low, and only about 4.5–13% and 1–3%, respectively, of the total number of enrolled learners obtain a certificate at the end of their courses. Still, most research concentrates on completion, ignoring the certification problem, and especially its financial aspects. Thus, the research described in the present thesis aimed to investigate *paid certification in MOOCs, for the first time, in a comprehensive way, and as early as the first week of the course*, by exploring its various levels. First, the latent correlation between learner activities and their paid certification decisions was examined by (1) statistically comparing the activities of non-paying learners with course purchasers and (2) predicting paid certification using different machine learning (ML) techniques. Our temporal (weekly) analysis showed statistical significance at various levels when comparing the activities of non-paying learners with those of the certificate purchasers across the five courses analysed. Furthermore, we used the learner's activities (number of step accesses, attempts, correct and wrong answers, and time spent on learning steps) to build our paid certification predictor, which achieved promising balanced accuracies (BAs), ranging from 0.77 to 0.95. Having employed simple predictions based on a few clickstream variables, we then analysed more in-depth what other information can be extracted from MOOC interaction (namely discussion forums) for paid certification prediction. However, to better explore the learners' discussion forums, we built, as an *original contribution, MOOCSent, a cross-platform review-based sentiment classifier, using over 1.2 million MOOC sentiment-labelled reviews*. MOOCSent addresses various limitations of the current sentiment classifiers including (1) using one single source of data (previous literature on sentiment classification in MOOCs was based on single platforms only, and hence less generalisable, with relatively low number of instances compared to our obtained dataset;) (2) lower model outputs, where most of the current models are based on 2-polar classifier (positive or negative only); (3) disregarding important sentiment indicators, such as emojis and emoticons, during text embedding; and (4) reporting average performance metrics only, preventing the evaluation of model performance at the level of class (sentiment). Finally, and with the help of MOOCSent, we used the learners' discussion forums to predict paid certification after annotating learners' comments and replies with the sentiment using MOOCSent. This multi-input model contains raw data (learner textual inputs), sentiment classification generated by MOOCSent, computed features (number of likes received for each textual input), and several features extracted from the texts (character counts, word counts, and part of speech (POS) tags for each textual instance). This experiment adopted various deep predictive approaches – specifically that allow multi-input architecture - to early (i.e., weekly) investigate if data obtained from MOOC learners' interaction in discussion forums can predict learners' purchase decisions (certification). Considering the staggeringly low rate of paid certification in MOOCs, this present thesis contributes to the knowledge and field of MOOC learner analytics with predicting paid certification, for the first time, at such a comprehensive (with data from over 200 thousand learners from 5 different discipline courses), actionable (analysing learners decision from the first week of the course) and longitudinal (with 23 runs from 2013 to 2017) scale. The present thesis contributes with (1) investigating various conventional and deep ML approaches for predicting paid certification in MOOCs using learner clickstreams (Chapter 5) and course discussion forums (Chapter 7), (2) building the largest MOOC sentiment classifier (MOOCSent) based on learners' reviews of the courses from the leading MOOC platforms, namely Coursera, FutureLearn and Udemy, and handles emojis and emoticons using dedicated lexicons that contain over three thousand corresponding explanatory words/phrases, (3) proposing and developing, for the first time, multi-input model for predicting certification based on the data from discussion forums which synchronously processes the textual (comments and replies) and numerical (number of likes posted and received, sentiments) data from the forums, adapting the suitable classifier for each type of data as explained in detail in Chapter 7.

Declaration

The work in this thesis is based on research conducted in the Department of Computer Science at Durham University. I declare that all the work in this thesis was performed by the author except where otherwise indicated and that no part of this thesis has been submitted elsewhere for any degree or qualification.

List of Publications

Given below is a list of the works from the thesis that have already been published:

Chapter 5: This chapter contains two published studies – a journal article at the International Journal for Artificial Intelligence in Education (IJAIED) and a conference paper at the International Conference on Intelligent Tutoring Systems (ITS).

- Alshehri, M., Alamri, A., Cristea, A. I., & Stewart, C. D. (2021). Towards Designing Profitable Courses: Predicting Student Purchasing Behaviour in MOOCs. *International Journal of Artificial Intelligence in Education*, 31(2), 215-233.
- Alshehri, M., Alamri, A., & Cristea, A. I. (2021). Predicting Certification in MOOCs Based on Learners' Weekly Activities. *International Conference on Intelligent Tutoring Systems* (pp. 173-185). Springer, Cham.

Chapter 6: This chapter contains a published conference paper at the International Conference on Information Systems Development (ISD).

- Alshehri, M. A., Alrajhi, L. M., Alamri, A., & Cristea, A. I. (2021). MOOCSent: A Sentiment Predictor for Massive Open Online Courses. *29th International Conference on Information Systems Development*. Association for Information Systems (AIS).

Chapter 7: This chapter contains a published conference poster at the International Conference on Artificial Intelligence in Education (AIED).

- Alshehri, M. and Cristea, A. I. (2022). MOOCs Paid Certification Prediction Using Students Discussion Forums. In Artificial Intelligence in Education. Posters and Late Breaking

Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part II: Springer, 542-545.

Forthcoming Publications

Given below is a work from the thesis that is under publication:

Chapter 3: This chapter contains a journal article submitted to ACM Computing Surveys.

- Alshehri, M., & Cristea, A. I. (2022). Prediction of Certification in MOOCs: A Systematic Literature Review. ACM Computing Surveys. (under review)

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Alexandra Cristea, for her constant support throughout my study and research. She was there to help me whenever I needed it. I have benefited considerably from her broad expertise while writing this thesis and participating in several scientific conferences. In short, I could not have completed this thesis without her ceaseless advice and guidance.

My deepest thanks go to my lovely wife, Muna, for her constant encouragement. She has been a great source of inspiration to me throughout this journey. It means a great deal to me to have her by my side, despite the fact that she is busy with her own postgraduate course and raising our two children, Sahab and Abdullah.

I am also deeply grateful to my father, mother and siblings for their support and their belief in me. They have all been a source of motivation, even from over 4,000 miles away. I would also like to thank my lab mates, especially Ahmed Alamri, Zakaria Alhassan and Laila Alrajhi, for their invaluable advice regarding my research.

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

ANN Artificial Neural Network

DL Deep Learning

DT Decision Tree

ET Extra Tree

FFNN Feedforward Neural Network

GRU Gated Recurrent Units

LA Learner Analytics

LR Logistic Regression

ML Machine Learning

MOOCs Massive Open Online Courses

NB Naïve Bayes

NLP Natural Language Processing

POS Part of Speech

RF Random Forest

SA Sentiment Analysis

SMOTE Synthetic Minority Over-Sampling Technique

TF-IDF Term Frequency Inverse Document Frequency

Chapter 1 : Introduction

Online courses have been around for decades; however, they generally cater to a limited audience (Ng and Widom, 2014). To address this limitation and other e-learning challenges, massive open online courses (MOOCs) were developed specifically to reach an unlimited number of potential learners worldwide. MOOCs, within a concise history, have attracted a great deal of attention from many stakeholders across the entire educational landscape, primarily due to their easy accessibility, both in terms of physicality (online access) and cost (low or no cost) (Castillo et al., 2015; González Robinson, 2016), rendering them more inclusive than other forms of education (Longstaff, 2014).

Tracing their history from MIT's 2001 OpenCourseWare[1] initiative, the term MOOC was officially coined in 2008 (Moreno-Marcos et al., 2018b), inspired by massive multiplayer online role-playing games (MMORPGs) (Greene, Oswald and Pomerantz, 2015). Later, MOOCs entered the modern age of successful commercialisation with today's giant platforms, such as edX[2], Udacity[3], and Coursera[4] , in 2011 (Ng and Widom, 2014). The following year, 2012, highlighted unprecedented growth in audience and market and was named "The Year of the MOOC" (Pappano, 2012; Reich and Ruipérez-Valiente, 2019). Following the three pioneers above, many other platforms have been launched worldwide, such as

[1] www.ocw.mit.edu

[2] www.edx.org

[3] www.udacity.com

[4] www.coursera.org

FutureLearn[5] in the UK, MiriadaX[6] in Spain, iversity[7] in Germany, XuetangX[8] in China, Veduca[9] in Brazil, and Schoo[10] in Japan (Qiu et al., 2016).

MOOCs have become increasingly popular, and their scale and availability make it possible to offer a diverse set of (free or cheap) learning content to learners from all over the world in an accessible and engaging manner. With the low barrier of access, MOOCs have been gaining more registered learners annually after just over a decade. Thus, many MOOC providers, such as Coursera, edX, and FutureLearn – the top worldwide MOOC providers in terms of offerings – have started offering scalable online courses to the public.

By the end of 2021, the number of MOOCs had reached almost 20,000 courses delivered via more than 950 university partners worldwide. The total number of MOOC learners has surpassed 220 million (Shah, 2021a). In 2018, the number of MOOCs platforms was 31 (Jaganathan, Sugundan and Sivakumar, 2018), but our research observed that this number had doubled as of 2022 to 63 platforms launched from 25 countries; just over half of them (n = 32) were from four countries, namely the United States (n = 17), China (n = 6), Japan (n = 5), and Italy (n = 4). Figure 1.1 shows the distribution of these platforms by country of establishment. It is worth mentioning that some platforms are launched/owned through cross-continent initiatives. Examples include the European Multiple MOOC Aggregator (EMMA)[11] or even across the globe, FutureLearn, jointly owned by the Open University in the UK and the Australian human resource consulting company SEEK Ltd. Thus, we tagged these platforms to the country of establishment or the location of their current headquarters.

[5] www.futurelearn.com

[6] www.miriadax.net

[7] www.iversity.org

[8] www.xuetangx.com

[9] www.veduca.org

[10] www.schoo.jp

[11] www.platform.europeanmoocs.eu

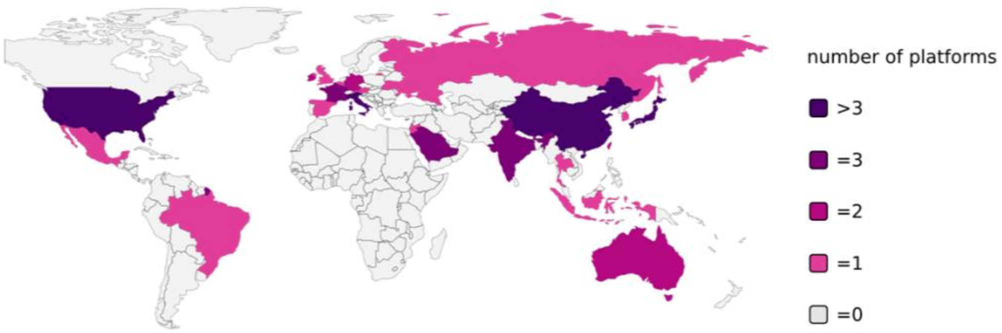


Figure 1.1. Global distribution of MOOC platforms. Locations denote the country of establishment of the platforms or their current headquarters.

Compared to traditional online courses, MOOCs are open to any potential learner and typically charge fees from certificate-earning learners only[12]. When they were sponsored by the business incubators of top-ranked universities, MOOCs were initially offered for free, at least to audit. This offering was concurrent with a global initiative to democratise education, which has helped attract a substantial number of registered learners (Dillahunt, Wang and Teasley, 2014; Lohr, 2020). This status marks the first few years of the emergence of MOOCs – the initial era of free and open MOOCs (Zhu, Sari and Lee, 2020). Later, with these platforms becoming more independent educational companies, monetised content has been necessary to fund the MOOCs and ensure the sustainability of these platforms. This trend resulted in monetised standard courses and included new forms of monetised content, such as micro-credentials, corporate training, and university degrees (Shcherbinin, Kruchinin and Ivanov, 2019; Cobos and Olmos, 2018).

1.1. Research Problem

Keeping the recent proliferation of MOOCs in mind, there are some indications that the number of registered learners and course populations are declining due to the transition of these platforms from semi-free to paywalled courses (Chuang and Ho, 2016). However, despite the unparalleled success of MOOCs, especially in terms of the burgeoning learner enrolment, one of the more disturbing aspects to date is the

[12] This applies neither to the free-certificate eligible courses (e.g. government initiatives) nor to the prepaid (paywalled) courses (e.g. degrees and corporate training courses).

staggeringly decreasing certification rates (Reich and Ruipérez-Valiente, 2019), a funnel with learners "leaking out" at various points along the learning pathway (Clow, 2013; Breslow et al., 2013). More critical is the issue that the paid certification percentage of a given course has also been declining over various runs/iterations; for instance, in some cases the number of certificate purchases dropped by as much as 50% in the latest course run compared to the first run (Alshehri, Alamri and Cristea, 2021). This challenging and constantly low certification rate has prompted substantial research on the topic (Gitinabard et al., 2018) and pushed several providers to explore potential business models for increasing revenues via numerous promotive strategies (Dellarocas and Van Alstyne, 2013). These strategies included, for instance, freely releasing a particular portion of the course content and requiring fees only to earn verified certificates and credentials (Reich and Ruipérez-Valiente, 2019). While this strategy has not yet contributed to improving the paid certification rate, the platforms still need a sustainable and sufficient source of revenue to recoup their operational costs.

1.2. Research Motivations

MOOC platforms have partnered with higher education institutions (HEIs), governmental bodies, and private corporations worldwide to provide certified academic curriculums and professional training. This offers an incredible opportunity to learners, considering the flexibility of MOOCs in terms of openness and accessibility from anywhere, facilitating a near-university campus learning experience (Almatrafi and Johri, 2018). It also helps MOOCs have enrolment arrays in thousands from diverse learners; nevertheless, the certification level within these courses is low. Therefore, addressing this challenge, which threatens platform sustainability, is a principal source of inspiration to conduct the existing research.

Another motive is the lack of certification predictive models, which marks only 9% (*n* = 9/78) (Gardner and Brooks, 2018b) to 14% (*n* = 14/94) (Moreno-Marcos et al., 2018b) of the outputs of the total current MOOCs predictive models. This dearth is to some extent surprising and uninterpretable, considering the current inevitable transition of the platforms into paywalled commercialised content was intended to initially meet their operational costs.

Machine learning (ML) techniques have been widely adopted not only in the educational setting (Kučak, Juričić and Đambić, 2018) in general but also more specifically in MOOC analytics (Zhu, Sari and Lee, 2020; Dalipi, Imran and Kastrati, 2018), addressing several challenges. With their incredible ability to handle big data within prediction and classification tasks, such techniques seem essential in analysing data-rich tasks such as MOOC dataset-based prediction tasks. With this in mind, the researcher was motivated to apply various ML techniques (based on the type of data under analysis as further explained in Chapter 5, 6, and 7) to examine whether learner data can help predict course certificate attainment.

Ultimately, this thesis is motivated by the desire to accurately detect learners not attaining their certificates, i.e., non-paying learners, as early as possible, to allow platform owners to intercede early and provide learners with any personalised intervention needed, which may in turn positively affect the certification rate (Rohloff, Sauer and Meinel, 2020), i.e. may convince the learner to purchase a certificate and, subsequently, help platforms build a more sustainable business model (King and Lee, 2022).

1.3. Research Questions

This thesis employs several ML models to investigate whether MOOC learners' data can help predict their certification attainment at an early stage of the course. The research questions addressed in this thesis were developed based on the research gaps and the limitations of the current MOOC certification predictive models (explained in detail at the end of Chapter 3). Therefore, the umbrella research question this thesis tackles is: ***How can learners' data in MOOCs be utilised for predicting paid certification?***

To help answer this broad research question, the following sub-questions were formulated:

- *RQ1: Do non-paying MOOC learners behave differently from course purchasers as to their activities of access and answering questions (attempts, correct and wrong answers)?*

This research question aims to statistically compare non-paying learners' activities versus certificate purchasers' activities to measure the extent to which the two groups' activities differ. Subsequently, the second research question examines whether learners' activities can be used to predict later certification behaviour.

- *RQ2: Can MOOC learner's clickstream data (accesses, attempts, correct and wrong answers) and time spent on course steps predict paid certification for courses?*

Considering clickstream is a very rich source of data in MOOCs, we employed learners' raw clickstreams (step accesses, correct and wrong answers) and computed time spent by each learner on each step (learning unit) to examine if they can predict learner certification attainment (or, more specifically, course certificate purchasing) in Chapter 5.

After analysing simple predictions based on a few clickstream variables, we then analyse more in-depth what other information can be extracted from MOOC interaction (namely discussion forums) for certification prediction. However, before exploring the learners' discussion forums, the second experiment (in Chapter 6) contains the procedure followed in building MOOCSent – a cross-platform, review-based sentiment classifier – using over 1.2 million MOOC reviews to train the model. Since the text dataset used in this specific experiment is unlabelled with learners' sentiments, which are ideal determinants of learner success in MOOCs (Sraidi *et al.*, 2022; Wen, Yang and Rose, 2014; Chaplot, Rhim and Kim, 2015; Dalipi, Zdravkova and Ahlgren, 2021), the pre-step is utilising MOOC discussion forums for predicting certification is labelling learners' posts (comments and replies) with the corresponding sentiments which would help enrich our training dataset and improve the performance of the predictive model (as further discussed in section 6.3).

Thus the third sub-question of this thesis was framed as follows:

- *RQ3: Can course reviews obtained from multiple MOOC platforms be used to build a reliable sentiment classifier?*

In the final experiment (in Chapter 7), learners discussion forums' raw data (learner textual inputs), sentiment classification using MOOCSent, and computed features (number of likes received for each textual input), in addition to several features extracted from the texts (e.g. character counts, word counts, and part of speech (POS) tags for each textual instance) were used to predict paid certification. Thus, the fourth research question this thesis addresses is as follows:

- *RQ4: Can raw and computed data extracted from MOOC discussion forums predict paid certification for courses?*

1.4. Research Objectives

This research project aims to examine the ability of ML to predict MOOC purchase (certification) at an early stage of the courses. To address the identified research questions in Section 1.3, the following research objectives were considered during this project:

- **RO1**: To first survey the current MOOC certification predictive models, synthesise the results for a comprehensive and deep understanding of this field, elucidate the limitations of these models, and propose some areas of improvement for future works. This is an essential preliminary step for addressing the main (umbrella) question of this thesis. Further explanation is presented in Chapter 3.

- **RO2**: To statistically examine the difference in behaviour between non-paying learners and certificate earners to examine the extent to which the two groups behave differently. Research question 1 is addressed by this objective in Chapter 5.
- **RO3**: To assess the capability of ML approaches for predicting certification in MOOCs using learners' clickstream data (Chapter 5). Research question 2 is addressed by this objective.
- **RO4**: To examine the extent to which the state-of-the-art NLP models can predict learners' sentiments based on their end-of-course reviews. This is the main objective that can be achieved by addressing RQ3. Further details are presented in Chapter 6.
- **RO5**: To assess to the capability of ML approaches for predicting certification in MOOCs using data from MOOC discussion forums (Chapter 7). Research question 4 is addressed by this objective.
- **RO6**: To identify and adopt the most suitable approach for dealing with imbalanced data during building ML predictive models, considering the highly imbalanced data utilised in this research project. This can be done by adopting class-weighting predictive models (Chapter 5), augmenting the minor class data with a text augmenting model for textual data (Chapter 6), or generating new synthetic instances of the existing minority cases for numerical data (Chapter 7).

1.5. Thesis Contributions

This thesis contributes to the knowledge and field of MOOC learner analytics with the following:

- Systematically reviewing the literature using the 27-item checklist of Preferred Reporting Items for Systematic Review and Meta-analysis (PRISMA) Protocol for methodological rigour to increase the transparency and quality of the literature synthesis. The present review reveals several limitations within the works surveyed that were addressed later during this research project and presented an insight into the current state of the art in certification predictive modelling. Further details are discussed in Chapter 3 and 8.

- Collecting a 5-course rich learners’ dataset from FutureLearn that – although it is the world’s third largest platform in terms of the number of courses offered (Shah, 2021a) – has never been utilised for modelling and predicting paid certification according to our Systematic Literature Review (SLR) in Chapter 3.

- Investigating various conventional and deep ML approaches for predicting paid certification in MOOCs using learner clickstreams (Chapter 5) and course discussion forums (Chapter 7).

- Building the largest MOOC sentiment classifier (MOOCSent) based on learners’ reviews of the courses from the leading MOOC platforms, namely Coursera, FutureLearn and Udemy, which handles emojis and emoticons using dedicated lexicons that contain over three thousand corresponding explanatory words/phrases.

- Developing a novel multi-input model for predicting certification based on the data from discussion forums. This model synchronously processes the textual (comments and replies) and numerical (number of likes posted and received, sentiments) data from the forums, adapting the suitable classifier for each type of data as explained in detail in Chapter 7.

- Enhancing the understanding of learner behaviour through identifying the factors (features) associated with learners’ decision to purchase a course certificate (as explained in sections 5.3.4 and 5.3.5). This can assist course providers, educators, and policymakers design effective strategies to promote and optimise the uptake of paid certification options.

- Developing *early* predictive models - using various types of learners’ data - that can anticipate learners' decisions regarding paid certification in MOOCs at an early stage of the course. These models can be employed by course providers for timely forecasting learners' likelihood of opting for paid certification, enabling them to allocate resources more efficiently, tailor marketing efforts, and optimise course design.

- Providing the potential for MOOC platforms for personalised interventions and support mechanisms for learners, taking into account their classified categories (non-paying or certificate purchasing). Accordingly, course providers can proactively engage with these learners, offer targeted incentives or assistance, and address any concerns they may have. This personalised approach is expected to enhance learner satisfaction, increase certification rates, and contribute to the overall success of MOOC platforms.

- Offering valuable insights for strategic decision-making and market analysis in the MOOC industry. Platforms can leverage the findings and predictive models to assess the viability of introducing new paid certification options, modify existing pricing structures, or target specific learner segments. This enables providers to make data-driven decisions, adapt to evolving market dynamics, and align their offerings with learner preferences and demands.

1.6. Thesis Outline

The research problem, scope, motivations, questions, objectives, and contributions have been outlined above, and the remainder of this thesis is organised as follows:

- Chapter 2 (MOOCs: A Business Perspective): This chapter introduces a brief definition and history of MOOCs and presents the different types of MOOCs; it also provides a description of various MOOC platforms. Additionally, MOOC business models and, more specifically, monetisation tiers are discussed.
- Chapter 3 (Systematic Literature Review): This chapter reviews the current MOOC certification predictive models in a systematic way using the PRISMA protocol. The chapter concludes with an organised synthesis of the works surveyed and also elaborates further on related topics such as the state of the art, limitations of the current models, and future opportunities for development.
- Chapter 4 (Methodology): This chapter describes the methodology followed for answering the research questions addressed in this thesis. This includes the data used, any preprocessing steps conducted, the feature selection techniques, and the classification algorithms employed. Additionally, this chapter explains other experiment-related matters such as model evaluation (performance) metrics and visualisation tools adopted in the modelling experiments.
- Chapter 5 (Predicting Paid Certification in MOOCs using Learners’ Weekly Clickstreams): This chapter contains an experiment on using the data on learners’ access, question answering, attempts, and time spent on each learning unit to predict paid certification (certificate purchase). The chapter describes the collected data and the preprocessing steps implemented. Next, the utilised approach and employed learning algorithms are discussed. Finally, the evaluation processes and results are presented and discussed.

- Chapter 6 (MOOCSent: A Sentiment Predictor for Massive Open Online Courses): This chapter compares the performance of the most common SA techniques using the largest collected MOOC review dataset. Firstly, the data collected and used for this experiment is described. Then, the utilised learning algorithms and the experiment setting are discussed. This is followed by reporting the model performance (per algorithm) and discussing the obtained results.

- Chapter 7 (Discussion Forum-based Prediction of Paid Certification in MOOCs): This chapter presents an experiment using learners' interaction in the discussion forums to predict paid certification. Firstly, the methodology followed for preprocessing textual and numerical data is presented. Next, the experiment setting is explained. Finally, the models' performance metrics and results are presented and discussed.

- Chapter 8 (Discussion): This chapter contains an extensive discussion of the thesis regarding its achieved results and novelty. It also discusses how these results add to the knowledge considering the existing literature. Additionally, it explains how the gaps identified within the SLR were addressed in the present thesis.

- Chapter 9 (Conclusion): This chapter summarises the key contributions of the thesis and discusses the opportunities for future development. Figure 1.2 illustrates the thesis outline workflow with the key contributions in bold maroon.

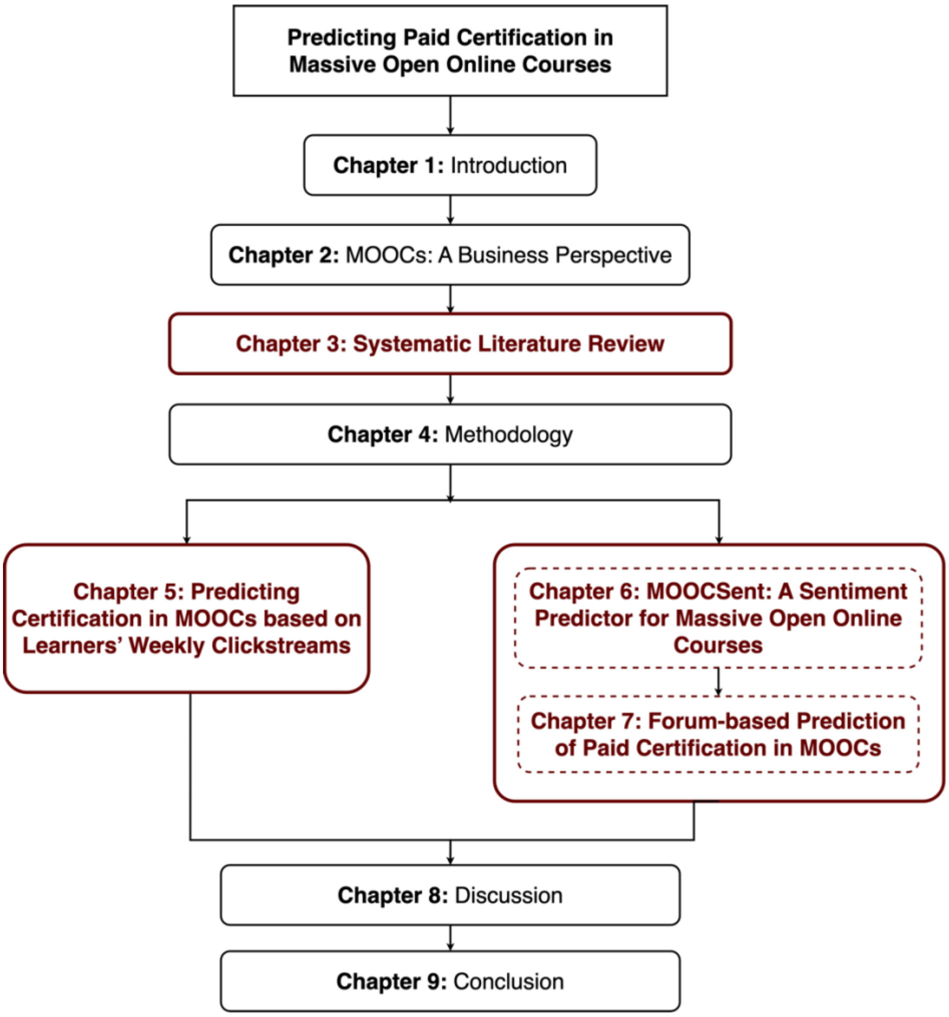

Predicting Paid Certification in Massive Open Online Courses
Chapter 1: Introduction
Chapter 2: MOOCs: A Business Perspective
Chapter 3: Systematic Literature Review
Chapter 4: Methodology
Chapter 5: Predicting Certification in MOOCs based on Learners' Weekly Clickstreams
Chapter 6: MOOCSent: A Sentiment Predictor for Massive Open Online Courses
Chapter 7: Forum-based Prediction of Paid Certification in MOOCs
Chapter 8: Discussion
Chapter 9: Conclusion

Figure 1.2. Thesis outline workflow with the contributions in bold maroon.

Chapter 2 : MOOCs: A Business Perspective

2.1. Prologue

This thesis is about MOOC Paid Certification Prediction. To better understand the context of our umbrella research question, we need to explain its context: that of MOOCs, but also, that of the paid certification, and where it sits within the business side of MOOCs. Hence, this chapter introduces a brief definition and history of MOOCs and the various types of MOOCs currently in use; it also provides a description of the MOOC platforms. Additionally, MOOC business models and, more specifically, monetisation tiers are discussed.

2.2. Definition of MOOCs

MOOCs are a new form of distance learning defined as "institutionally-based formal education where the learning group is separated and where interactive communication technologies are used to connect the instructor, learners, and resources" (Simonson, Zvacek and Smaldino, 2019). Nevertheless, this definition includes the earlier version of the MOOCs business model, where they were institutionally based and embraced by the world's leading universities, such as edX (founded by MIT and Harvard), Coursera, and

Udacity of Stanford University, FutureLearn of Open University, and XuentangX of Chinese Tsinghua University (Notaris, 2019; Lohr, 2020). The above platforms and the majority of the other subsequently launched platforms benefit from Higher Education Institutions (HEIs) as business accelerators during their first forming stages and Another difference between MOOCs and distance learning is instructors' involvement in the teaching and learning processes. With the massive number of participants, a MOOC instructor can typically be the MOOC designer or the talent featured in a video displayed at any time point of the course without real-time human intervention. In addition, MOOC content is typically a web-based digitised representation (e.g. quizzes are self-marked and exams are programmed with self-scoring). Therefore, a distance learning course usually contains two essential components of the education process, namely distance teaching AND distance learning, which are not usually present in MOOCs (Simonson, Zvacek and Smaldino, 2019).

Based on the above characteristics, MOOCs appear different (in terms of teaching) and more flexible (in terms of learning) at various levels compared to traditional distance learning courses. Thus, a MOOC can be defined as "an online course designed for a large number of participants that can be accessed by almost anyone anywhere, as long as they have an internet connection, is open to everyone without entry qualifications and offers a full/complete course experience online for free" (Brouns et al., 2014). This definition is more precise; nevertheless, it neglects the recent monetisation strategy that many platforms have followed. A more recent and financially aware definition of MOOCs is "an online learning environment that learners have open access to and can register for free or with low cost" (Zhu, Sari and Lee, 2020). Today, MOOCs are freely open to any potential learner, with a lower barrier to access, allowing them to gain millions of learners annually, and typically charge fees from certificate-earning learners only (Zhu, Sari and Lee, 2020).

In this thesis we have used mostly the FutureLearn MOOC environment, due to convenience sampling, as explained in Section 4.2. We have, however, used also other platforms, for broader coverage of the various MOOC types, such as Coursera, Udemy, FutureLearn, Stanford in Chapter 6.

2.3. Defining MOOC Certification

MOOCs have been attracting more learners over time; nevertheless, statistics show that there has been a decrease in the number of yearly certificate earners due to the transition towards paid content (Cagiltay, Cagiltay and Celik, 2020). With the platforms monetising their courses, this new business model for MOOCs, coinciding with repeated iterations (course reruns) of the same content, has caused a noticeable decline in the uptake of many courses over subsequent runs (Chuang and Ho, 2016). Moreover, even if the enrolment figures seem massive, the low certification rate, of just around 13% of the enrolled learners, and the worse paid certification rate, of less than 1% of the enrolled learners, appear as real challenges, which attracts the attention of many previous predictive works on MOOCs (Jordan, 2014; Alshehri *et al.*, 2021; Alshehri, Alamri and Cristea, 2021; Alsheri *et al.*, 2021; Arslan, Bagchi and Ryu, 2015). It is worth mentioning here that paywalled courses (e.g. university degrees) have better certification statistics, and 40–90% of the enrolled learners earn a certificate of completion at the end of the course (Lohr, 2020).

Considering MOOCs have their own characteristics, adopting traditional educational metrics such as certification to this unique form of education renders prediction in MOOCs a more challenging task (DeBoer et al., 2014). Additionally, while MOOCs are proliferating alongside the continuously emerging private corporation-managed platforms, a clear-cut and comprehensive definition of MOOCs is becoming more challenging. Thus, there is a need to define MOOC certification more clearly.

Certification, compared to other MOOC outcomes, such as success, dropout, or learning achievements, which are pedagogically complex to measure or controversial to determine (Cobos and Jurado, 2018; Gitinabard et al., 2018), seems quite straightforward to define. Platforms set particular requirements (e.g. completing specific steps and scoring above a threshold for assignments and quizzes) to make the learner eligible for the course certificate (Blackmon and Major, 2016). While course completion and certification may sometimes be used interchangeably, they have different determinants and neither pre-require the other. A MOOC learner can complete a course without earning a certificate, especially in courses that do not require tuition fees in advance. Also, a learner can earn a certificate even before completing the course but after meeting the minimum requirements set by the platform. Completion typically refers to completing the entire course, usually defined by a specific number of steps in the form of videos, assignments, and quizzes. However, a literature review showed that previous studies have a different definition of completion (Gardner and Brooks, 2018b). Certification in MOOCs can be defined as earning enough points on course assignments to meet some predetermined threshold for the course (typically around 70%) and earning a certificate of accomplishment based on that (Gardner and Brooks, 2018b). However, this definition neglected the financial aspect of the certification, where most MOOCs require paying a predefined fee for obtaining a certificate of completion. As opposed to this, paid certification refers to earning a certificate that, unlike completion, has its own payment-related factors (Ruipérez-Valiente et al., 2017).

2.4. MOOCs: A Unique Concept of E-learning

Recent technological advancements have revolutionised the educational domain and blessed it with many learning experience improvements (Raja and Nagasubramani, 2018). MOOC platforms can virtually deliver their content worldwide among these new forms of education. They are unique in their characteristics compared to other forms of technology-based education. Their *massiveness* allows a much more extensive number of learners than in traditional or technology-enhanced classrooms. The number of enrolled learners reaches hundreds for specialised courses and thousands for more generic courses (Bonafini, 2018; Buholzer, Rietsche and Söllner, 2018). It is believed that MOOCs have emerged in response to the pressure put on universities to offer education that is accessible to all at any stage of their lives. Since their emergence, they have shown an overwhelming learner–teacher ratio compared to traditional e-learning classes. Besides the promotional free access to their learning content, this helps MOOCs attract significant attention from both the media and the educational community (Atiaja and Proenza, 2016).

Openness plays a significant role in MOOC expansion, where learners can join a massive array of courses delivered by, as of 2021, almost one thousand universities worldwide (Shah, 2021a). Most courses have no prerequisites compared to traditional learning, and course materials on MOOC platforms can be accessed smoothly via the internet by diverse learners in terms of demographics, level of education, location, and profession. This encourages the heterogeneousness of learner characteristics and the intentions of populations in MOOCs (Koller et al., 2013; Chuang and Ho, 2016). Although MOOC populations in some platforms skew towards specific characteristics in terms of country of origin, gender, and level of education (Greene, Oswald and Pomerantz, 2015), they are still significantly more diverse than any other traditional or e-learning environments (Glass, Shiokawa-Baklan and Saltarelli, 2016; Christensen et al., 2013). Nevertheless, identifying these highly diverse demographics and intentions (i.e. learner characteristics) is challenging. Although most MOOCs provide pre-course questionnaires to their learners, they are usually not mandatory to enrol in a given course so as to lower the entry barriers. This has led, however, to a shallow response rate to learner demographics questionnaires (Kizilcec and Halawa, 2015; Whitehill et al., 2015; DeBoer et al., 2013). Consequently, most works on MOOC predictive modelling have focused on using learner's activities, rather than their demographics, for modelling different MOOCs outputs (e.g. completion, dropout, or certification) (Gardner and Brooks, 2018b) (Moreno-Marcos et al., 2018b).

The lax nature of completion and certification is another unique feature of MOOCs. In traditional face-to-face and e-learning courses, learner performance (e.g. quizzes and exam results) is strictly considered a critical factor in counting academic credits or awarding official certificates after the end of the course. MOOCs, in contrast, more conveniently allow learners to reattempt automatically marked quizzes or even retake the whole course in future runs (iterations) without any penalty imposed for previous failures or delays in completing a task. This level of *freedom* makes MOOCs a suitable learning environment for learners who want to learn a specific skill without the need to attend the whole course, hence having little or no interest in completion and, consequently, certificate attainment (Borrego, 2019; Zhu, 2021).

Although some MOOCs have predefined start and end dates and intermediate due dates and are generally divided into time-boxed weekly learning content and introduced to learners on timely releases (Calise et al., 2019; Wang, Hemberg and O'Reilly, 2019), their *asynchronicity* in attendance is considered unique. Learners, especially in computer-graded MOOCs, do not have to "stick" to a specific timetable to attend the courses and interact either with the course content (e.g. videos) or with other learners (e.g. by posting in the course forum) (Mullaney and Reich, 2015). This high level of self-paced flexibility allows learners to freely view course content and interactively participate with others at any time.

Whilst we are not directly targeting the learning side in this thesis, we include this for completeness, as learning is the primary target for MOOCs, which has to be taken into account even when predicting other factors (like Paid Certification).

2.5. MOOCs History

MOOCs have been gaining more interest since their introduction due to their early commitment to openness, promising access to worldwide top-ranking university education. Many early MOOC initiatives were funded by the Hewlett Foundation as part of their Open Educational Resources (OER) portfolio, aiming to provide openly accessible educational content to learners worldwide, including MIT Open Courseware and the Open University's OpenLearn (Casserly, 2018). Nevertheless, nowadays, the concept of MOOCs has radically changed with the increasing emergence of MOOCs monetisation and investors' willingness to return on their investments (Belleflamme and Jacqmin, 2016).

The first MOOC was the University of Manitoba's Connectivism and Connective Knowledge course (CCK08), which attracted just over 2,000 learners (Fini, 2009), after which the acronym MOOC was coined (Yousef et al., 2014), although many of these can be traced back to earlier (specifically 2007) e-learning experiments such as the Introduction to Open Education, Social Media & Open Education and the OER movement in general (Iiyoshi and Kumar, 2010).

Considering the few enrollees in the above initiatives, the real launch of MOOCs, as something considerably massive, was with Stanford University's introduction of three courses in late 2011, where each course had around 100,000 enrolled learners from over 190 countries (Perez-Pena, 2012). Soon after in the same year, MIT's Open Online Learning initiative called MITx was announced. Later in 2012, Udacity was launched as a purely commercial enterprise (i.e. offering paywalled content only since its inception (Yousef and Sumner, 2021)), followed by the establishment of Coursera with a USD16 million fund from four major US universities. Hence, 2012 was the birth year of many of today's leading platforms, including MIT and Harvard's edX, with a fund of USD60 million from both institutions. The proliferation of MOOCs continued in 2013, with platforms launched outside the US for the first time. Two more leading platforms were launched that year: FutureLearn was introduced by 12 leading UK universities as a for-profit platform (Brown, 2013), and MiriadaX was launched by more than one thousand Latin American universities, offering courses for Spanish-speaking learners (Atiaja and Proenza, 2016).

After that many platforms were launched globally; a complete list of own elaboration of the present platforms is available in Appendix A. Figure 2.1 illustrates a timeline of the formation of MOOC platforms over the years and how business models have been updated. It shows how platforms have focused on paid content such as micro-credentials, corporate training, and degrees rather than aiming to deliver free courses. Over the last five years, free certificates or statements of accomplishment disappeared from most leading platforms, in line with the introduction of more paid content, such as corporate training and online degrees (Condé and Cisel, 2019). This is a logical transformation as platforms have belatedly realised free courses-only business model may not help in building a sustainable business model (Sharples, 2019)

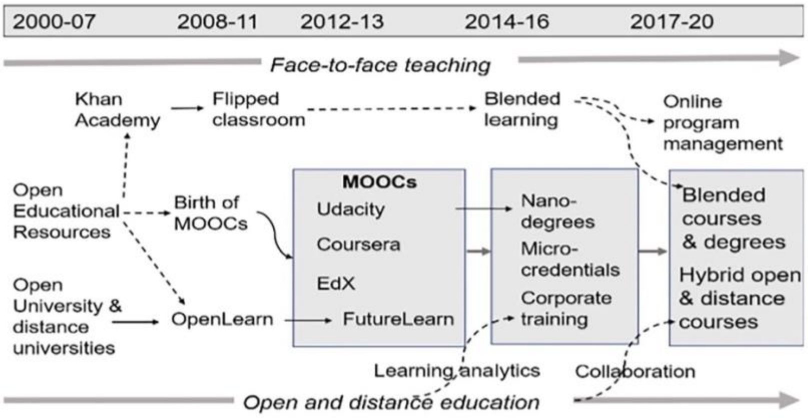


Figure 2.1. A timeline of MOOC platforms establishment (dashed arrow = influence; solid arrow = direct relation), cited from (Sharples, 2019).

In this thesis, the focus is mainly on the FutureLearn platform, which is the main European one, due to both its importance in the UK as well as convenience sampling. We have, however, used also other platforms, for broader coverage of the various MOOC types, such as Coursera, Udemy, FutureLearn, Stanford in Chapter 6.

2.6. Types of MOOCs

In contrast to the earlier 2007-2008 courses, the later MOOCs of 2011 onwards were different in content, assessment, and communication as they were not centred around networking and learner autonomy. They instead adopted a more traditional behaviourist strategy for offering learning content in consequent small units together with assessments based on multiple-choice tests to enable learners to monitor their own performance (Brown, 2013). This form of course design has become known as xMOOCs, while the earlier versions, which were learner centred, creative, and autonomous, were dubbed as cMOOCs because of their emphasis on connectivism (Hill, 2012).

As shown in Figure 2.2, cMOOCs are decentralised (autonomous), allowing learners to set appropriate learning objectives and content independently; thus, no formal curriculum is supplied because learners are involved in constructing the curriculum. cMOOCs are based on the connectivist theories and emphasise connecting learners rather than offering learning content (Borrás-Gené, 2019). As cMOOCs do not assess learners' progress, certificates of completion are not offered (Henukh et al., 2019). xMOOCs, in contrast, adopt the traditional university learning method where learning objectives are pre-defined by the instructor through short videos, lecturers, and instructor-led demonstrations (Yousef et al., 2014).

Communication and collaboration are essential differences between cMOOCs and xMOOCs. Learners are expected to build a shared understanding of the topic and discuss their learning outcomes with others in cMOOCs. The learning resources in this type of MOOCs are a collection of outside-the-platform content such as Google Docs, Blogs, and other leading social media platforms such as YouTube, Twitter, and Facebook. xMOOCs, in contrast, host the learning content, such as videos and provide integrated discussion forums where learners can interact with peers unlimitedly, in addition to the ability to set notifications for any updates on the course (Daniel, 2012). Regarding certification, cMOOCs do not grant credentials to learners being self-organised and assessed, whereas xMOOCs, by design, take pre-steps for certificate eligibility, including computer-based assessments (Koutropoulos, 2013).

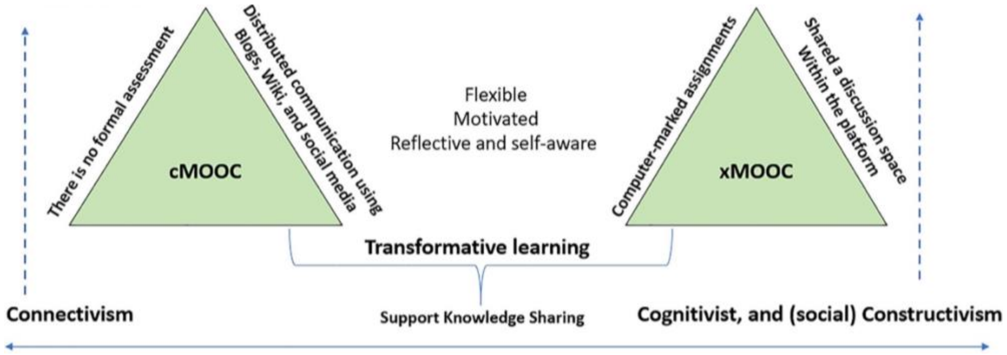

There is no formal assessment
Distributed communication using Blogs, Wiki, and social media
cMOOC
Flexible
Motivated
Reflective and self-aware
Computer-marked assignments
Shared a discussion space Within the platform
xMOOC
Transformative learning
Support Knowledge Sharing
Connectivism
Cognitivist, and (social) Constructivism

Figure 2.2. Characteristics of cMOOCs versus xMOOCs, cited from (Yousef and Sumner, 2021).

In this thesis, the focus is on xMOOCs, where certification is part of the process, as this is what the aim of the prediction is.

2.7. MOOCs as of 2022

The last couple of years, coinciding with the spread of COVID-19 and the consequential lockdown of educational institutions and travel restrictions in many countries, were exceptional in MOOC history. About one-third of the total learners (60 million) joined in 2020 (Shah, 2020). By 2020, MOOCs had exceeded 180 million learners. The number of offerings reached 16,300 courses, of which 2,800 were launched in 2020. Additionally, the number of offered micro-credentials in 2020 increased substantially, with 360 new micro-credentials, from 820 in 2019. Furthermore, 19 new worldwide online degrees were introduced (a one-quarter increase) in 2020, totalling 67 MOOC-based degrees. This massive amount of content has been delivered by around 950 universities worldwide (Shah, 2020). One instance of the MOOC boom during the pandemic is Coursera, which had 10 million newly enrolled learners from mid-March to mid-May in 2020, which was 7 times the pace of new enrolment in the previous year. New enrolments at other counterparts, such as edX and Udacity, have also increased by similar multiples (Lohr, 2020). The following year showed continuously increasing 2020-like statistics. In 2021, MOOCs reached 220 million learners, 19,400 courses, 1,670 micro-credentials, and 70 MOOC-based degrees (Shah, 2021a). The proliferation of MOOCs continued, and the number of platforms reached 63 in 2022.

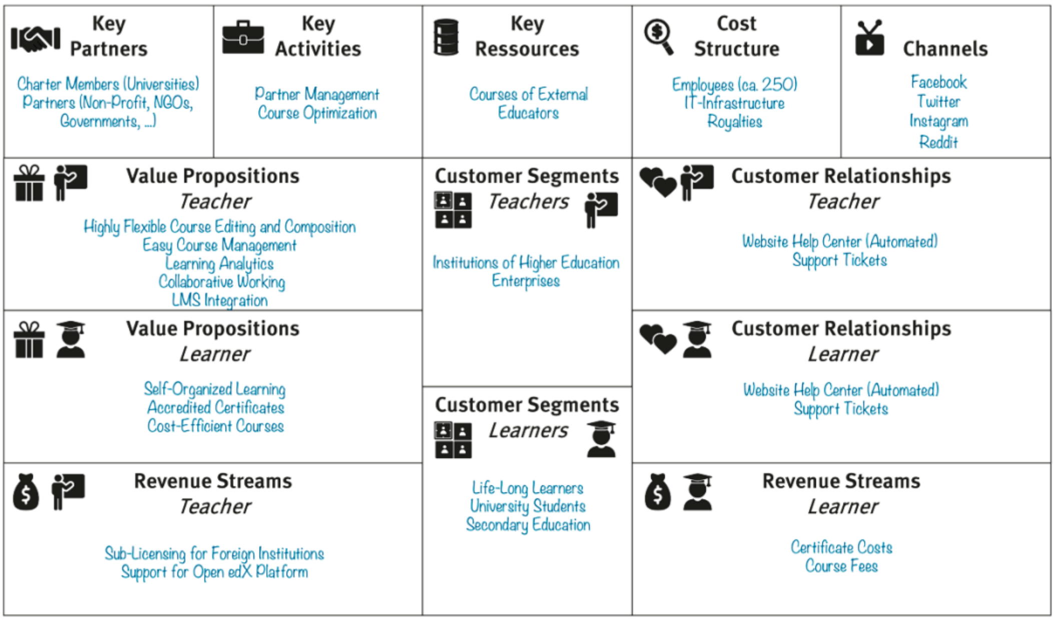
Such statistics show the importance of this research, in terms of primarily monetary implications on the MOOC providers, and secondarily, on the educational reach of the MOOCs.

2.8. MOOC Business Models

The initial aim of MOOCs was to provide access to educational resources to vulnerable learners. With this concept, MOOCs were expected to disrupt the education sector by “democratising learning” offering free courses at an economic scale via recording the course once and iteratively offering it to millions of learners (McGreal et al., 2013). However, it has been challenging to recover the substantial costs of running a platform, where many technical and human resources are involved, with free or even low-cost standard courses only (Halsbenning and Niemann, 2021). While revenue generation has been a primary challenge (Burd, Smith and Reisman, 2015), platforms are expected to adopt a sustainable business model to meet their economic validity (Cusumano, 2013).

The business models of universal electronic platforms have been found inappropriate for MOOC platforms because (1) the platforms are still in an immature stage with frequently-changeable business functionality (Farrow, 2019); (2) the complexity of investment in education where the value and return on investment (ROI) can only be determined post hoc; and (3) the competitive presence of government-funded educational initiatives, offering content for free and making the survival of MOOCs more critical (Halsbenning and Niemann, 2021). Hence, bespoke business models to consider the potential success opportunities and threats are essential for the sustainability of these platforms.

Figure 2.3 shows a generic business model derived from most platforms (35 platforms) with exemplificatory notes of edX business model elements (Halsbenning and Niemann, 2021). The business model was built using the highly adopted business model canvas (BMC) with fine adjustments for MOOC platforms. BMC is defined as “a strategic management tool that provides a visual representation of a company's business model, serving as a framework for describing, designing, and analyzing a business model (Osterwalder and Pigneur, 2010). Key Partners include charter members (e.g. universities), governments and non-governmental organisations (NGOs). Instructors and learners were split into customer relations and value propositions due to being substantially different customer types, requiring different service levels and support. Instructors and learners were also split into customer segments as platform content may be offered for institutional learners such as universities and corporations rather than individual learners.



Key Partners

Charter Members (Universities)
Partners (Non-Profit, NGOs, Governments, ...)

Key Activities

Partner Management
Course Optimization

Key Ressources

Courses of External Educators

Cost Structure

Employees (ca. 250)
IT-Infrastructure
Royalties

Channels

Facebook
Twitter
Instagram
Reddit

Value Propositions *Teacher*

Highly Flexible Course Editing and Composition
Easy Course Management
Learning Analytics
Collaborative Working
LMS Integration

Customer Segments *Teachers*

Institutions of Higher Education
Enterprises

Customer Relationships *Teacher*

Website Help Center (Automated)
Support Tickets

Value Propositions *Learner*

Self-Organized Learning
Accredited Certificates
Cost-Efficient Courses

Customer Segments *Learners*

Life-Long Learners
University Students
Secondary Education

Customer Relationships *Learner*

Website Help Center (Automated)
Support Tickets

Revenue Streams *Teacher*

Sub-Licensing for Foreign Institutions
Support for Open edX Platform

Revenue Streams *Learner*

Certificate Costs
Course Fees

Figure 2.3. Business model of MOOC platforms with exemplificatory notes.

Revenue streams (learners) include various or even opposing monetisation strategies. For example, some platforms do not allow access to content before paying the fee, even in standard courses. In contrast, other platforms do not rely on any obligatory payments by learners, granting access to the whole course content and only charging for certificate issuance, preceded by achieving the minimum course-specific grade. This indicates that while MOOCs are generally offered similarly to the consequent, usually weekly, learning units (Brown, 2013), each platform has its own strategy for content monetisation, which is further explored in the next section.

2.9. Monetisation of MOOCs

Initially, MOOCs were developed to “democratise education” by offering free access to learners who cannot afford formal education (Dillahunt, Wang and Teasley, 2014; Lohr, 2020). With this proclaimed mission, the early courses attracted thousands of learners from different geographical and cultural backgrounds (Nkuyubwatsi, 2014). Nevertheless, these platforms later started seeking funds and forming separate learning corporations due to initial sponsorship streams drying out (Paldy, 2013). For instance, while edX and Udacity were founded as institution-based nonprofit organisations, with all their offering free in 2012, they later had Silicon Valley's leading venture firms funding these platforms to form their own business models, following the classic internet formula of "lure a big audience and figure out a business model later" (Lohr, 2020). During the early years of the emergence of MOOCs, platforms did not have a transparent partner-wise business model and kept exploring potential revenue opportunities with the assistance of their HIE partners (Baturay, 2015). This included, but was not limited to, several monetisation scenarios: (1) charging university partners for platform technical support, (2) signing profit-sharing agreements with partners, and (3) charging course hosting fees based on the length of the course (Milheim, 2013). The evolutionary history MOOCs, from a financial perspective, can be divided into two phases:

- 2009-2016, when courses were primarily offered free of charge.
- 2017 to the present time, when most providers began monetising content, stopped offering free certification, and consequently generated revenues by offering credentials (Zhu, Sari and Lee, 2020) (Condé and Cisel, 2019).

With this new business model, providers, besides the classic single type of content (e.g. single time-limited courses), have begun developing additional educational products that range in price from low cost to highly expensive to enrol in, for example, degrees from world-class universities) (Wang, Hemberg and O'Reilly, 2019; Shah, 2018b). While the standard courses offered by universities are still critical products on MOOC platforms, further additions, such as credentials by corporate and government entities, have been introduced. Hence, the essential "free for enrolment" open to any potential learner on MOOCs can act as a marketing principle for attracting more course certificate purchasers.

In their journey to adopting the new financial business model, MOOC providers have gone (or are going) through six tiers of MOOC monetisation, as illustrated in Figure 2.4 (Shah, 2018b). These process steps range from the earlier free standalone courses to the recent fully paywalled online degrees. However, while this process of categorisation seems to be generally followed by many platforms, it has not necessarily been adopted by every platform towards monetising content (Taneja and Goel, 2014).

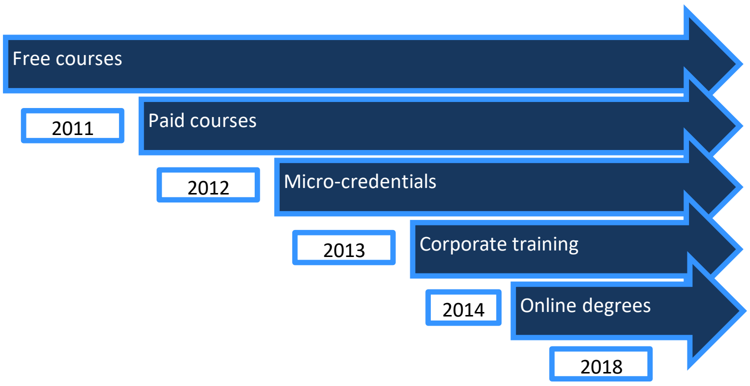


Figure 2.4. Tiers of MOOC monetisation, dates denote the first introduction of each tier, inspired by Shah (2018b).

2.9.1. Free Courses

The first tier is where a learner can audit a single course free of charge, typically without earning a certificate of completion. This is the most common style of learning in MOOCs, where the overwhelming majority of learners tend to learn for free. While most MOOC platforms offer free course auditing, only charging for certification (Liyanagunawardena et al., 2019), some courses are offered free, including a free certificate at the end of the course. However, these courses are scarce, accounting for less than 3% of MOOCs on some platforms as of 2021 (331 out of 10,004 on Coursera[13] and 42 of 1,582 on FutureLearn[14]) and are usually sponsored to be offered for a free certificate. Additionally, some platforms offer learners the choice to pay for the certificates of some courses. For example, edX allows the learner to receive a free certificate "honour code" or pay a fee for a "verified certificate"[15].

The freemium offering, which is the most common MOOC pricing strategy (Porter, 2015), includes linking a sequence of services/products to the learner, where the first (i.e. course auditing) is offered at no cost and the latter (i.e. course certificate), which extends the free service, is offered at a certain fee. This service/product marketing strategy is prevalent in promoting MOOCs (Baker and Passmore, 2016).

[13] https://www.coursera.org/search

[14] https://www.futurelearn.com/courses

[15] https://edx.readthedocs.io/projects/open-edx-learner-guide/en/named-release-cypress/SFD_certificates.html#:~:text=Honor%20code%20certificates%20are%20free,requirements%20to%20pass%20the%20course.

2.9.2. Certified Courses

The following phase is where monetisation begins upgrading to obtain a course certificate of completion, still at a pre-defined low cost, ranging from a few tens of dollars to the low hundreds of dollars (Cagiltay, Cagiltay and Celik, 2020; Goli, Chintagunta and Sriram, 2019). This model is based on learners wishing to provide evidence of upskilling to potential employees. It requires learners to achieve a grade (e.g. 60% or more in edX and over 70% in FutureLearn for certificate purchasing eligibility) (Ruipérez-Valiente *et al.*, 2017; Lee, 2018b). They are the lowest cost compared to the other tiers discussed below, offering the lowest threshold for a potentially more significant population. One of the main features of certified courses, compared to the following tiers, which typically require further instructor intervention, is the automated learning process on the provider platforms. Lectures in such courses are pre-recorded, and assignments are auto-graded and peer-reviewed[16.]

Additionally, learner contributions in the discussion forums can be automatically analysed and marked based on their relevance to the course topic (García-Molina et al., 2020). Regarding paid certified courses, some platforms, such as Coursera[17], offer financial aid or scholarships for learners who cannot afford course fees. Also, Coursera further categorises certificates based on the obtained score, where learners who achieve 65% or more of the maximum possible score receive a standard certificate, and those who achieve 85% or more of the maximum possible score receive a distinction certificate (Jiang, Fitzhugh and Warschauer, 2014).

2.9.3. Micro-credentials

The third tier of MOOCs monetisation is represented by micro-credentials, a non-degree credentialing paradigm that consists of multiple single courses (typically spreading over six months in length). These range in price from a few hundred dollars to a few thousand dollars based on the programme content and the provider platform (Shah, 2018b; Pickard, Shah and De Simone, 2018). Micro-credentials are multi-course series that have been offered since 2013 (with the edX introduction of XSeries) as a midpoint between standalone courses and fully online degrees (Shah, 2021b) and still offer relatively affordable and accessible learning content (Lemoine and Richardson, 2015). Again, depending on the offering platform,

[16] https://www.coursera.org/browse

[17] https://www.coursera.support/s/article/209819033-Apply-for-Financial-Aid-or-a-Scholarship?language=en_US

these courses have different trademark names (e.g. Specialisations[18] on Coursera, XSeries[19] on edX, ExpertTracks[20] on FutureLearn, and Nanodegrees[21] on Udacity) (Pickard, Shah and De Simone, 2018).

The primary purpose of micro-credential development is to signal proficiency or experience in a specific area, which is less broad than an online degree covers. Although micro-credentials are usually not regulated and accredited by certifying third parties such as online degrees offered by universities, their role in reaching short-term goals is significant. This includes becoming qualified for a particular career or closing a skill gap in a specific area (Krauss, 2017; Lewis and Lodge, 2016). In addition, some micro-credentials, especially longer and more expensive ones, offer credit earning for certain online degrees. Thus, learners who have already attended a specific micro-credential will be awarded credits for specific online degrees. This academic crediting scheme is available on some leading platforms, such as Coursera and edX (Pickard, Shah and De Simone, 2018). This is a valid business model, gradually bringing customers from cheaper offers to more substantial ones.

The micro-credential pricing strategies vary by platform. For example, while some platforms require the full fee in advance, others allow learners to pay for each course individually. Another charging method is subscription-based learning, where a pre-defined tuition fee is payable on a monthly or term basis, such as Specialisations in Coursera and Nanodegrees on Udacity (Pickard, Shah and De Simone, 2018).

2.9.4. Corporate Training

The fourth tier of monetisation (corporate training) has targeted audiences from minor teams to the most prominent organisations since its introduction in 2014 (Bogdan et al., 2017). MOOCs have not only succeeded in academia but also in professional development, fostering the training of employees in a flexible way (Cobos and Olmos, 2018), for instance, the staff training agreement between Udacity and the IT service management company Pearson VUE and the agreement between MiriadaX and the Spanish telecommunication company Telefonica (White et al., 2014). The cost varies and can be on a per-user-per-year basis or based on a custom pricing agreement, especially with large organisations and government entities (Shah, 2018b). MOOC providers have extensively adopted professional training over the past few years (Calise et al., 2019). At the same time, this has attracted the attention of both the private and public

[18] https://www.coursera.org/specialization
[19] https://www.edx.org/xseries
[20] https://www.futurelearn.com/experttracks
[21] https://www.udacity.com/nanodegree

sectors for a more flexible approach to staff training and upskilling (Notaris, 2019). With MOOC-based corporate training, the geographical constraints on staff training have been left behind.

Additionally, training courses can now be monitored interactively due to the privileges the platforms grant to the corporates throughout, facilitating the administrative and legal affairs of the training courses. This helps monitor an employee's progress over the training sessions and identify any course-related issues in real time (Condé and Cisel, 2019). Besides academic success, MOOCs have been used to foster employee development and provide required training (Vivian, Falkner and Falkner, 2014). Professional learning, which MOOCs typically support, is critical for developing and maintaining expertise in today's workplace (Milligan and Littlejohn, 2014). Financially, corporate training has been adopted by MOOC platforms to monetise content and increase revenues (Arslan, Bagchi and Ryu, 2015).

2.9.5. Online Degrees

MOOCs were promoted for offering free courses as a promising solution to the global demand for higher education. Recently, platforms have been integrated into higher education institutions to develop paid programmes, including blended and fully online degree programmes. The offered courses are designed for post-secondary education and early and mid-career professionals who want to master specific job-related skills (Littenberg-Tobias and Reich, 2020).

This last monetisation tier is where higher educational institutes are deeply involved in offering learners a campus-like learning experience. This includes further services (e.g. mentorship, office hours, and supervised exams). MOOC-based online degrees cost a few thousand to tens of thousands of dollars (Shah, 2018b) and include bachelor's and master's degrees, mainly in technology, science, and business. The MOOC master's degree takes two to three years and costs from USD16,000–22,000. Beyond the online accessibility from anywhere around the world, their attraction is that they are slightly cheaper than the traditional version of the same courses while promising similar quality (Shcherbinin, Kruchinin and Ivanov, 2019).

2.10. Trends and Implications Resulting from MOOC Business Models

With this emergent phenomenon of a new monetisation scheme, providers have been able to transfer their free courses into revenue-generating educational content. This new development responds to the need for MOOC platforms, with their university partners, to reach a sustainable revenue model. However, this new orientation has affected MOOC statistics in recent years (Zhu, Sari and Lee, 2020). Several new trends that can be ascribed to MOOC monetisation have been observed, such as (1) the decrease in the number of annual newly enrolled learners; (2) the increasing offering of college credits, credentials, and degrees (Hollands and Kazi, 2019; Shah, 2018a); and (3) the prevalence of business-oriented services (e.g. courses dedicated to upskilling employees) (Schaffhauser, 2018; Shah, 2019).

With the recent transition of MOOC platforms towards monetising their content, focusing on courses that generate more revenues (Lohr, 2020), a fully asynchronous new business model has been introduced that allows learners to access the entire course content on demand and complete the course at their own pace (Gardner and Brooks, 2018b). This transition seems to be an essential survival plan for many more platforms. For example, Udacity was almost forced out of business in 2018, laying off around half its workforce. However, the platform has survived with a further focus on corporate training (especially employee reskilling) and paid content in general (Lohr, 2020).

The above distinctive features of MOOC financing outline an unmatched learning environment that is highly unique compared to other extensively studied forms of learning (e.g. on-campus and e-learning). Of the various business models implemented, the ones based on certification are the oldest and, thus, the longest running. Unlike the other forms of monetisation, which often demand (at least partial) upfront financial commitment in the form of payment models based on certification, they are highly risky, as they expect learners to pay at the end of their studies. However, they can also bring in potentially higher gains, as their low cost opens them to a much wider learner population than any of the other financial models. Even a small percentage of a substantial learner body [e.g. mid-career professionals (Littenberg-Tobias and Reich, 2020) and female learners with higher degrees (Samuelsen and Khalil, 2018)] could bring in potentially higher revenues than other streams. Thus, a mechanism to understand when and how they can transition to paid certification is essential. Certification, among other profitable services of MOOCs, for example, course hosting fees and linking learners to potential employers, seems to be among the highest revenue drivers (Brown, 2013). Therefore, it is vital to survey previous certification predictive models. While MOOCs were analysed in the past and reviewed in the literature, including in surveys, **there is no literature review up to now synthesising and analysing previous studies on MOOC certification predictive modelling**; this is the gap we are addressing in this study.

2.11. Epilogue

This chapter introduces a definition of MOOCs and MOOC certification. It also discusses the history and the types and platforms of MOOCs. Additionally, the data sources for modelling, business models, and the various tiers of monetisation are also discussed. The following chapter contains the methodology, synthesis, and results of our literature survey concerning certification prediction in MOOCs.

Chapter 3 : Systematic Literature Review

3.1. Prologue

This chapter reviews the current MOOC certification predictive models in a systematic way using the PRISMA protocol. The chapter concludes with an organised synthesis of the works surveyed along with further elaboration on related topics such as the limitations of the current models and future opportunities for development.

3.2. Introduction

Considering the modernness of MOOCs, being around for a decade and that providing platforms are still at their critical stage of building solid business models (Pappano, 2012), a clear review of previous certification models would help shed light on how these models have dealt with the low certification rate, i.e., predicted certification across the different MOOCs platforms. The current SLR, which deals with a total of 25 works predicting certification in MOOCs, as explained in detail in Section 0, is considered necessary to evaluate these studies' definitions of MOOC certification and findings, as well as compare the results based on the methodologies the surveyed works followed. This objective is essential, considering the different definitions of certification, i.e., free (where the prediction task does not need to include, typically, a learner's financial decisions) or paid (where the learner pays a certain fee to earn a certificate of completion). An explicit exploration of how the surveyed models are similar/different based on the type of certification being addressed is essential.

While most platforms generally follow similar approaches in terms of content delivery, looking in-depth into their designed courses, the method of teaching, the type of learning content or even the data aggregated from different platforms show that each platform has its unique method of certification. Thus, a predictive model/algorithm may not be suitable for different courses/platforms (Lee, 2018b). For instance, some platforms provide learners with more visuals compared to audio content, others allow individuals to create and deliver courses, whereas some other platforms restrict this to institution-affiliated educators only. Similarly, the payment protocol for the course certificate differs, where it can be a pre-requirement to join the course or done after finishing the whole. This logically stands against the concept of a one-fits-all predictive model, where each platform, dataset, sample of learners or prediction task has its own characteristics (Rizvi et al., 2022). Therefore, the conducted SLR compares the current MOOCs certification predictive models, based on all of the above factors, for a better insight into the state-of-art in MOOCs certification prediction.

3.3. Previous Surveys on MOOCs

Previous MOOC-related reviews of the literature have addressed several pedagogical concerns. These include some theoretical reviews, such as modelling learning and assessment in MOOCs (Joksimović et al., 2018), self-regulated learning (Wong et al., 2019) and teaching and learning progress (Deng, Benckendorff and Gannaway, 2019). Other reviews considered the practical side of MOOC studies, such as research methods, topics, and trends of empirical MOOC research (Zhu, Sari and Lee, 2020), didactic applications for Foreign Language Learning (Palacios Hidalgo, Huertas Abril and Gómez Parra, 2020) and the role of motivation in retention (Badali et al., 2022). Prediction-wise, a few studies have focused on the traditional challenges of low completion (e.g., predicting learner success (Gardner and Brooks, 2018b), predicting dropout (Mehrabi, Safarpour and Keshtkar, 2020; Dalipi, Imran and Kastrati, 2018) and identifying predictive model outcomes, features, and techniques used to evaluate predictive model performance) (Moreno-Marcos et al., 2018b). However, synthesising previous studies on MOOC certification, which include certification prediction models for free and paywalled courses, understanding MOOC business models, and exploring recent platform offerings (e.g., university degrees and corporate training), has not been done to the best of our knowledge. This coincides with the few studies on the understanding, analysis, and modelling of MOOC certification in comparison to other more studied sides of MOOCs mentioned above (Cagiltay, Cagiltay and Celik, 2020). Since the beginning of their unprecedented proliferation a decade ago, we believe MOOCs have gone through advancement, especially content monetisation, as discussed in detail in Section 2.9, which deserves a separate survey to explore these financial trends.

As the present study aims at surveying the previous works on MOOC certification predictive models since the emergence of MOOCs in 2011 (Ng and Widom, 2014) up until the end of 2021, we intend our study to be as inclusive as possible, keeping in mind the need to exclude any irrelevant previous work, which does not fall into our inclusion criteria as explained in Section 3.4.3. Further details on the inclusion and exclusion strategy are presented in Section 3.4.1.

3.4. Surveyed Resources

Our surveyed resources[22] include two typically used collection databases: Scopus[23] and Web of Science (WoS)[24], the two most comprehensive abstract and citation bibliographic databases of peer-reviewed scientific journals and conference proceedings, with more than three billion cited references combined (Web of Science, Confident research begins here. ; Pranckutė, 2021; Zhu and Liu, 2020). In addition to their tremendous number of indexed references, these databases index the majority of the publishers in the field of e-learning, and EDM, such as the Institute of Electrical and Electronic Engineers (IEEE) Xplore[25], Association for Computing Machinery (ACM) [26], Springer[27], Taylor & Francis Group[28], ELSEVIER[29] and ERIC[30]. Our adopted resources above index the conferences and journals, ordered alphabetically below, which are the typical venues for MOOC predictive modelling publications:

[22] Google Scholar was dropped from this survey for not allowing searching for keywords within abstracts, using wildcards and exporting the retrieved results.

[23] https://www.scopus.com

[24] https://www.webofscience.com

[25] https://ieeexplore.ieee.org

[26] https://www.acm.org

[27] https://www.springer.com

[28] https://www.tandfonline.com

[29] https://www.elsevier.com/en-gb

[30] https://eric.ed.gov

- British Journal of Educational Technology (BJET)[31].
- International Conference on Artificial Intelligence in Education (AIED)[32].
- International Conference on Educational Data Mining (EDM)[33].
- International Conference on Learning Analytics and Knowledge (LAK)[34].
- International Conference on Learning at Scale (L@S)[35].
- International Journal of Artificial Intelligence in Education (IJAIED)[36].
- Journal of Learning Analytics (JLA)[37].
- Journal of Educational Data Mining (JEDM)[38].

3.4.1. Inclusion and Exclusion Criteria

This survey includes journal articles and conference papers that meet some essential requirements: (1) being written in English, (2) peer-reviewed to ensure the highest standards of research rigour and credibility, (3) providing an adequate explanation of the data used, the feature engineering approach followed, the learning algorithms adopted, and the results achieved. Our survey excludes other publications, such as pre-print, book chapters, books, and magazines.

The keywords *mooc**, "*massive open online course**" and *Certif** were used for retrieving the surveyed works. The surveyed websites' search queries are not case-insensitive (i.e., the search keywords' different cases, such as 'MOOCs' or 'moocs', are treated alike). Keyword phrases, signalling the necessity for retrieving documents that include exact word order, such as 'Massive Open Online Course', were applied using quotation marks; this has helped filter out many unrelated search results. 'Predict*' was not included with the search keywords (1) to make the retrieved studies as comprehensive as possible (2) and avoid

[31] https://www.bera.ac.uk/publication/british-journal-of-education-technology

[32] https://iaied.org

[33] https://educationaldatamining.org

[34] https://dl.acm.org/conference/lak

[35] https://learningatscale.acm.org

[36] https://iaied.org/journal

[37] https://learning-analytics.info/index.php/JLA

[38] https://jedm.educationaldatamining.org/index.php/JEDM

missing any related studies tagged with synonyms (e.g., ‘classification’, ‘detection’ or ‘forecasting’), either within the title (Yeomans, Reich and Acm, 2017) or the abstract (Wang and Wang, 2019; Liao et al., 2017; Elbadrawy et al., 2016).

Wildcards such as the asterisk (*), interpreted as a substitute for any number of letters, were used. We used an asterisk after each root form of a search term to include either any potential popular form of our search keywords (such as ‘MOOC’ as well as ‘MOOCs’) or any other possible form of that search term (e.g., ‘certifi*’ also collects information containing ‘certificate’ or ‘certified’), allowing all terms beginning with the same root word to be included in the search. Additionally, we use parentheses to override the order of precedence, where the expression(s) inside the parentheses is/are executed first. These search techniques are typically standard across the publishers’ database websites (Scopus[39] and WoS[40]) surveyed(ACM Advanced Search; Springer Link Search Tips; IEEE Explore Search Tips; Web of Science Core Collection: Search Tips; Scpus: Tips and Tricks).

The database search using the above protocol retrieved 446 studies (WoS n = 250 and Scopus n = 196). After merging the two files, 114 duplicated studies were removed, resulting in 332 unique studies, as shown in Table 3.1. Journal articles represented just over half of the total studies (*n* = 170), whereas conference papers (*n* = 134) represented about 40% of the retrieved studies.

Table 3.1. The number of retrieved works from WoS and Scopus, distributed by the type of the work.

Ref. Type	**WoS**	**Scopus**
Journal Articles	88	82
Conference Papers	91	43
Book Chapters	11	7
Reviews	2	3
Early Accesses	0	3
Editorials	2	0
Total	194	138

[39] https://blog.scopus.com/tips-and-tricks

[40] https://clarivate.libguides.com/woscc/searchtips

Figure 3.1 shows the types of retrieved works distributed by the number of authors, where the dotted lines denote the mean, and the solid lines denote the median. The number of authors generally correlates with the significance of the work (i.e., journal articles and conference papers tend to have more authors).

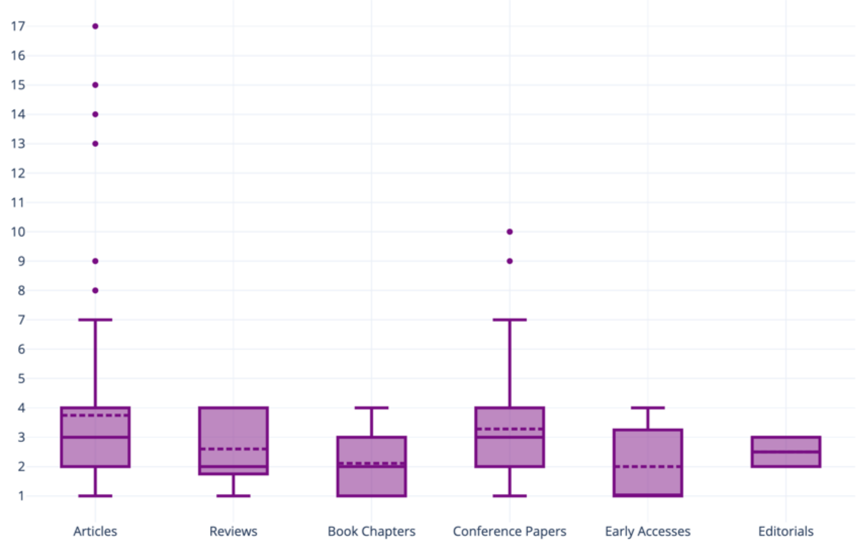


Figure 3.1. Types of retrieved studies distributed by the number of authors (M = solid lines, μ = dotted lines).

The left-hand geometric network in Figure 3.2 illustrates the unique studies (*n* = 332 nodes) and the authors' collaboration on different studies (157 edges). The edge weights represent the number of co-authors of multiple works. Regarding author collaboration in different works, the data show that nearly one-third (*n* = 108) of the studies were authored by at least one author of different studies. The right-hand figure illustrates a network of author collaborations, where nodes denote unique authors (*n* = 945), and edges denote co-authorship (*n* = 1930). The edge weights between authors here represent the number of co-authored works. There are only 31 sole authors who independently authored at least one study, whereas two-to-four-author studies were the highest among our retrieved studies, as shown below.

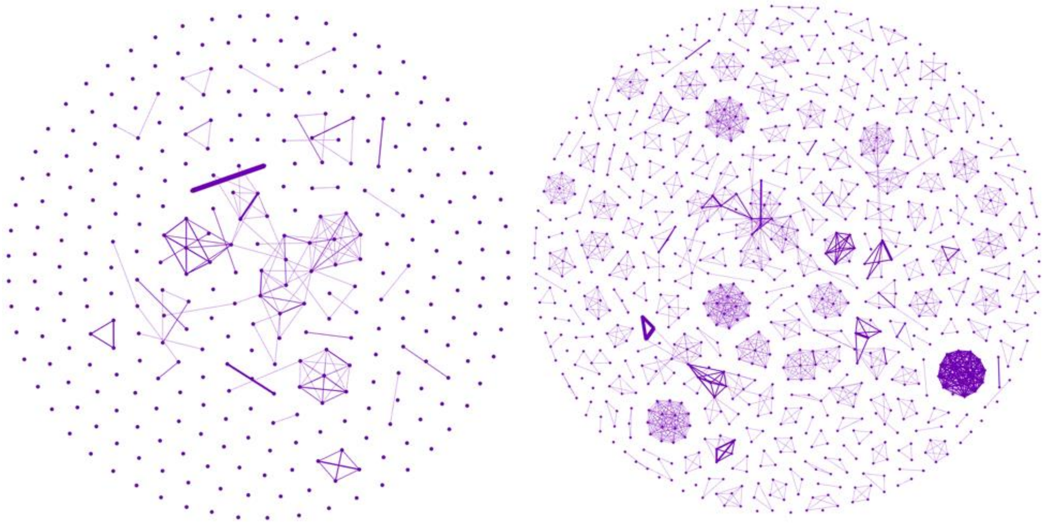


Figure 3.2. Study-to-study (left) versus author-to-author network (right).

3.4.2. Screening Process

For the screening step of the studies, we used Rayyan[41], a free, semi-automated online tool that expedites the screening of abstracts and titles in an interactive blind-reviewing environment (Ouzzani et al., 2016). Rayan is an AI-based tool that learns from preliminary raters' manual labelling. It provides labelling suggestions for the awaiting studies, reducing the load on raters and providing them with experimentally-approved high accuracy suggestions (Olofsson et al., 2017).

In this stage, three independent raters, all with a PhD degree and previous research publication experience, went through the 332 titles and abstracts to label them as included or excluded based on the criteria mentioned in Section 5.2 above. During the screening, we flexibly erred on the side of inclusion for studies that indirectly contributed to the literature on MOOC certification prediction. This includes statistical studies (e.g., studies that investigated the impact of course fee payment on learners' behaviours or the learners' demographic determinants of certification) (Goli, Chintagunta and Sriram, 2019; Arslan, Bagchi and Ryu, 2015) or studies that minorly predicted certification with, or as a subsequent target of other outcomes, e.g. dropout or grades (Xu and Yang, 2016). Figure 3.3 shows a snapshot of the summary

[41] https://rayyan.ai

report of the studies' screening process, which took almost 12 hours and 27 sessions on average for each rater.

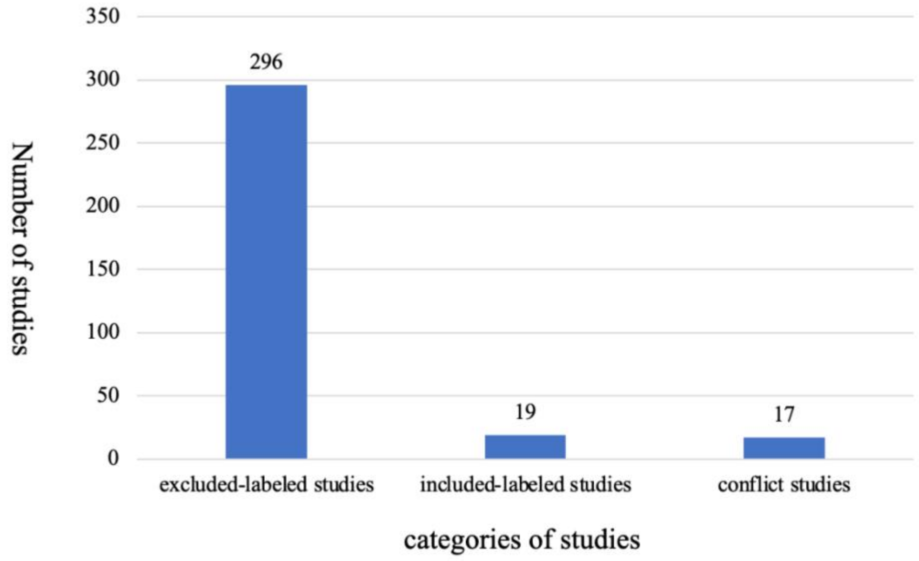


Figure 3.3. Summary of studies screening conducted by the three raters.

The screening resulted in 19 studies marked as included, 296 excluded, and 17 'conflict' studies. Rayyan optionally allows raters to attach their reasons for the inclusion/exclusion decision. Fleiss' Kappa, an adaptation of Cohen's kappa for assessing the reliability of agreement between three or more raters (McHugh, 2012), was used to measure interrater reliability. The test resulted in $k = 0.96$, which signifies a substantial agreement between raters (Fleiss, Levin and Paik, 1981).

Below is a list of the main exclusion reasons provided by raters, along with some instances of excluded studies:

- Different outcomes: studies that predicted a different outcome to certification (e.g., dropout, retention, and assignment grades) (Impey, Wenger and Austin, 2015; Haddadi and Dahmani, 2016; Rossano, Pesare and Roselli, 2017).
- Non-peer-reviewed studies: books, pre-prints, and editorials (Michael Spector, 2017; Zheng, Chen and Burgos, 2018b; Zheng, Chen and Burgos, 2018c).

- Publications in foreign languages other than English (Gardair et al., 2016; Garcia Barrera, Gomez Hernandez and Monge Lopez, 2017; Sánchez, 2016; Vrillon, 2019; Njingang Mbadjoin and Chaker, 2021).
- Assessment studies or surveys of MOOCs certification systems (Kumar, 2019; Kocdar, OKUR and Bozkurt, 2017; Fedorova and Skobleva, 2020; Canessa, Tenze and Salvatori, 2013).
- Studies that already present matching keywords but are actually on entirely unrelated topics (e.g., studies of MOOC cybersecurity certification systems and certificate authentication) (Beckerle, Chatzopoulou and Fischer-Hübner, 2021; Zheng, Chen and Burgos, 2018a).

3.4.3. Excluded Conflict Studies

A subsequent session to review the full manuscript of the conflict studies took place. The total number of conflict studies was 17, out of which 11 were excluded and six were included, rendering the total number of selected studies (*n* = 25). The excluded studies include Greene, Oswald and Pomerantz (2015), which, in contrast to the other works that target predicting certification, used survival analysis to examine the degree to which the learner's pre-stated intention of certificate attainment can predict the final exam result. The data analysed contains a pre-course survey within which a question about the learner's intention to obtain a course certificate after finishing the course. This feature, along with several other demographics and activities, was used to predict learners' final grades, which is the main requirement. Still, no actual certification was included in the data; hence, Mourdi et al. (2019) predicted learner success (course completion) and dropouts using a Stanford open edX dataset of around 3,500 learners. Lim, Tang and Ravichandran (2017) examined the mediating effects of learner intention for enrolment on the correlation between facilitating conditions and habit (independent variables) and the course's actual usage (dependent variable) by adapting the UTAUT2 model. Glance, Barrett and Hugh (2014) used log activities of 42 Stanford University MOOCs to examine the attrition rate within course auditors and active participants. Isidro, Carro and Ortigosa (2018) employed various shallow and deep learning techniques to predict dropout in MOOCs. The present study emphasised that predictive model performance does not necessarily positively correlate with complexity. The study showed that simple shallow algorithms (e.g., NB and Decision Tree) outperformed Long Short-term Memory (LSTM) in their dropout prediction task (Jiang, Zhang and Li, 2015) in Chinese.

All the above studies, while adhering to our search protocol’s keywords and having at least one included rating, do not fall within the outline of this SLR and are hence excluded. Figure 14 illustrates the PRISMA-based flow diagram of the study identification-to-selection process.

We followed the PRISMA framework (Moher *et al.*, 2009; Page *et al.*, 2021), the most frequently used and cited guideline for conducting systematic reviews and meta-analyses (Kite et al., 2015; Sitanggang et al., 2021; Page and Moher, 2017; O'Dea et al., 2021; Fleming, Koletsi and Pandis, 2014) to guide our research process. PRISMA contains four phases stepwise (identification, screening, eligibility, included) and a precise 27-item checklist (starting from how to title the present systematic review to declaring whether funding has been received to conduct the present systematic review) to increase the transparency and quality of the systematic review reported (Liberati et al., 2009). The protocol was followed while conducting this review; Nevertheless, some items were associated with meta analysis and thus not applicable to our analysis (see Appendix C for the full PRISMA 2020 checklist). Stage one involves developing the search protocol by determining the research questions, identifying the bibliographic databases, and defining the search keywords. Stages two and three subsequently involve applying inclusion and exclusion criteria. The last stage involved extracting data from the eligible studies and conducting the analysis, which was individually conducted by the author of the present thesis. Figure 3.4 illustrates this process in detail, along with the outcomes of each stage.

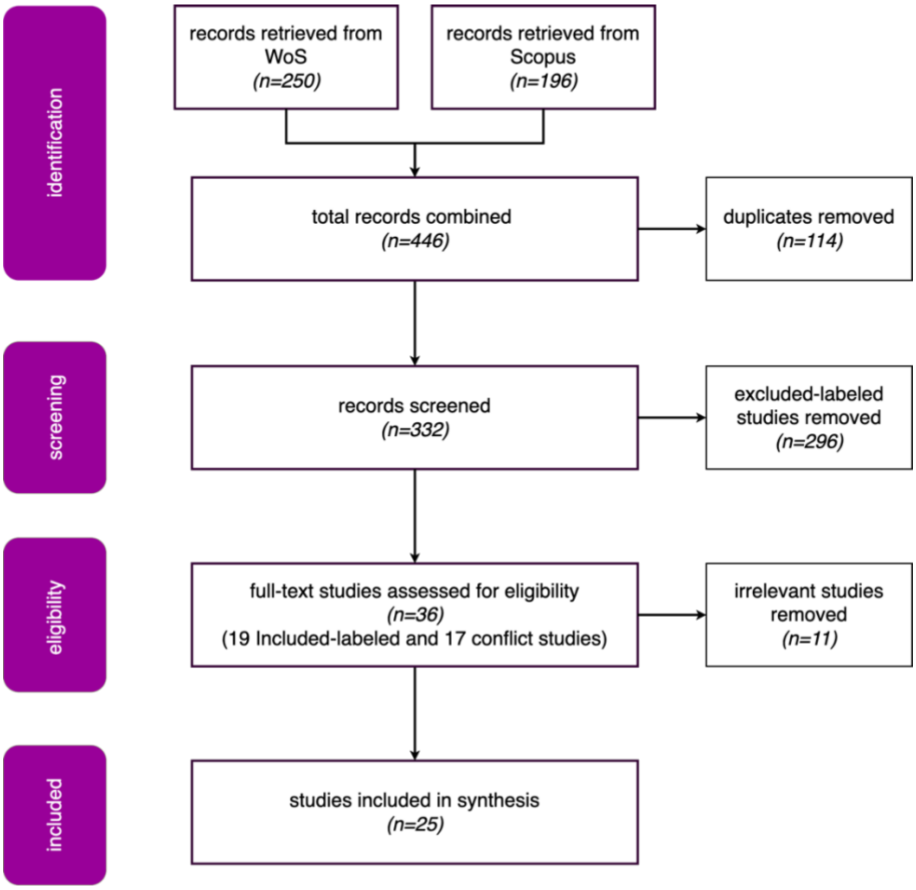

identification
screening
eligibility
included
records retrieved from WoS (n=250)
records retrieved from Scopus (n=196)
total records combined (n=446)
duplicates removed (n=114)
records screened (n=332)
excluded-labeled studies removed (n=296)
full-text studies assessed for eligibility (n=36) (19 Included-labeled and 17 conflict studies)
irrelevant studies removed (n=11)
studies included in synthesis (n=25)

Figure 3.4. PRISMA flow diagram

3.4.4. Categorisation Strategy

Although the works surveyed aim to predict certification in MOOCs, each experiment has its own characteristics. Thus, selected studies for inclusion cannot be directly compared based on the final prediction’s numerical results (e.g., performance metrics such as accuracy from ML models or p-value for statistical models). One of the reasons for this heterogeneousness is the study-by-study variation in the size of the population, the number of MOOCs analysed, the included runs of each course, and the type of data utilised. Additionally, the methodologies followed and the metrics reported were divergent across all the works surveyed. Thus, this study instead categorises the surveyed works mainly based on the general methodologies followed along with visual synthesis from different angles: the data sources (platforms), course delivery interval in years, size of the data utilised (number of learners, courses, and runs), types of the data (Clickstream, demographics, discussion forum-based), classification models/algorithms, performance metrics, prediction earliness and finally the type of certification (whether free or paid). For the visual representation of the synthesised studies, we used Plotly[42]: a Python graphing library which generates interactive web-based graphs (Sievert, 2020) as illustrated across Section 0.

3.5. Certification Prediction in MOOCs

3.5.1. Statistical Models

While the primary purpose of this study is to identify the previous works on predicting certification in MOOCs, we also included previous works that used statistical analysis to identify the determinants of certifications, compare ‘free’ learners versus certificate earners or works that measure the difference between the activities of both types of learners (free and certificate). The surveyed statistical models typically use course-level data to provide a general outline of certified learners’ characteristics and activities based on course metadata. The models of this orientation have been grouped based on the type of data used for analysis, as below.

3.5.1.1. Clickstream-based Models

Wintermute, Cisel and Lindner (2021) conducted a network-based exploration of learners’ achievements by examining how a course-course interaction affects the likelihood of certification. The certification rate is about 8% of registered learners. Using more than one million course registration events by almost 400 thousand learners on a French MOOC, the learners’ certification was modelled with Weibull Shape Parameter and Logistic Regression (LR). The study found that user engagement positively correlates with the certification rate in all 140 courses analysed. However, a registration burst (where a learner registers for multiple MOOCs within a short period) correlates negatively with the probability of certification. As suggested by the authors, this behaviour can be improved by restricting the registration to a pre-defined number of MOOCs over a certain period. However, this burst might represent a pattern of behaviours, such

[42] https://chart-studio.plotly.com/~shehri_m7#/

as the MOOC selection strategy, in which learners optimise their learning options, which may need further investigation.

Wang, Hemberg and O'Reilly (2019) studied the impact of learner-obtained grades on their activities (two groups of learners: certified learners and continuously participating learners) during the remaining content of the course using data from two edX MOOCs. The study found that the activity level positively correlates with the learner grade, and the delta activity variation (calculated from the difference in the activity before and after the finalisation of a grade) also correlates positively with the grade. Regarding certificate earners', delta grade (grade changes) was slightly different compared to delta activity, and active participants (who continually participated but were not interested in certification) had higher delta grades than certificate earners. It also found that learners' behaviours did not change significantly after reaching the minimal grade for certification and that the change in grades and activity is dependent on course characteristics (such as difficulty). More specifically, certified learners do not tend to change their behaviour during the course for later planned achievements. In particular, learners' grades should not be assigned precedence in significantly interpreting learner behaviours. While no solid relationship was found between grades and activities, low grades and dropouts correlated considerably.

To explore learners' intention–behaviour gap in MOOCs (Celik and Cagiltay, 2023) investigated the change in learners' intentions from completion and certification attainment to various outcomes. The study indicated that the intention-behaviour gap occurs in courses when learners do not reach the intended behaviours. The intention-behaviour gap for failing to act upon learners' positive intentions was caused by inclined abstainers. According to the findings, these abstainers were mainly related to the individual learner (e.g. lack of time, insufficient prior knowledge of the topic, taking another course from another platform), technical issues (issues related to connectivity, low computer specification) or course design (content is not straightforward or not as expected, late assignment grading, the need for more interactive courses). This study highlighted some concerns that MOOC providers should consider to reduce the attention behaviour gap.

3.5.1.2. Survey-based Models

Using a pre-course survey, Yeomans, Reich and Acm (2017) examined the impact of prompting MOOC participants' learning goals for success by reviewing learners' intention to earn a certificate, having achieved grades between 70% and 80%. The study found that prompting learners' pursuits in advance can increase the certification rate by 40%. The survey data revealed that almost 60% of learners intended to certify; nevertheless, only 16% obtained certificates at the end of the course. They used Latent Dirichlet Allocation (LDA) for clustering the survey-extracted textual data, showing promising results in forecasting certification. Learners who planned to certificate were more likely to explain how they would engage with the course than non-paying learners, who only specified when and where they would engage with the course content.

Using two runs of an introductory Python MOOC offered before and during the COVID-19 pandemic, Yee *et al.* (2022) examined how the pandemic influenced US learners' success in MOOCs. Revealing the correlation between various measures of COVID-19 severity and certification rate, the study found some of these relationships significant. The preliminary analysis showed that the pandemic led to a higher absolute number of enrollees and certificate earners; nevertheless, the certification rate dropped when local pandemic severity increased. While the pandemic resulted in more motives for enrolling on MOOCs, such as quarantining and increasing unemployment, learners may have lost their motivation due to the growing effect of COVID-19 in their locales. Additionally, a strong (negative) correlation between the pandemic new cases and the certification rate change within the two runs was observed. Since a deep understanding of this association was unavailable, further analysis of the causal mechanisms of pandemic-related stressors and certification statistics may help improve certification predictive modelling.

3.5.1.3. Discussion Forums-based Models

(Joksimović et al., 2016) conducted a social network analysis (SNA) based on learners' social interactions in the discussion forum to investigate the relationship between learners' social centrality measures (i.e., degree, closeness, and betweenness) and their certificate attainment. The study used descriptive and statistical SNA on two runs of the same course (one in English and one in Spanish). It concluded that learner structural centrality with reciprocal ties with peers (more interaction with other learners) positively correlates with the likelihood of certificate attainment at the end of the course. However, certificate earners in the course's English version (run) were more likely to interact in the forum.

Similarly, Jiang, Fitzhugh and Warschauer (2014) explored the association between learner centrality and performance using discussion forum data from two MOOCs but around double the number of learners analysed by . Learners were found to rarely interact with other learners in different performance groups (grades and attainments), suggesting that learners may use discussion forums to facilitate information flow and help-seeking rather than as a tool for social interaction with other learners. Additionally, the discussion forums were mainly used by a small percentage of learners who actively participated in commenting and replying, far more than their peers. Regarding certification, one of the MOOCs (algebra) showed that certified learners were more central in discussion forums. In contrast, the other (financial planning) showed no association between learner centrality and the likelihood of certificate attainment. According to the authors, one possible reason for the above difference might be the nature of the MOOC itself. Algebra is more academic and prepares learners to succeed in higher education, whereas financial planning is more of a life skills course. As a result, learners who were active in the discussion forums in the latter course may not have been concerned about certificate attainment.

Liu *et al.* (2022a) used discussion forums' emotional and cognitive engagement, which have an interactive relationship but are rarely analysed at a deep level, to forecast learning achievements in MOOCs. The developed text classifier aimed at automatically detecting emotional and cognitive engagement and identifying their sophisticated relationships with course outcomes. The developed model first used interpretable NLP techniques for recognising emotional and cognitive engagement patterns using data from 8867 learners' discussions. Next, the relationship between emotional and cognitive engagement and achievement was analysed. The structural equation modelling shows that learning achievement is highly influenced by learners' emotional and cognitive engagement, especially with confused and positive emotions, which were highly correlated with learning achievements compared to negative emotions. The findings also indicated that co-occurring emotion and cognition indicators were more reliable predictors of learning achievement than other variables. This study reveals the significant role of learners' emotions in discussion forums as a predictor of course outcomes and learning achievements.

3.5.1.4. Multi-sources-based Models

Some studies have used more than one type of data to model the certification in MOOCs. For instance, Arslan, Bagchi and Ryu (2015) explored the variables associated with MOOC certification using multi-level modelling: learner's characteristics, economy, culture-related variables, and country-level infrastructure of 24 developed and developing countries. The findings suggest that internet bandwidth and Hofstede's uncertainty avoidance cultural dimension are positively associated with the likelihood of earning a MOOC certificate. Demographically, the study further found that gender, age, and level of education significantly correlate with certificate attainment on edX. Additionally, determinants of certification differ between developed and developing countries. Country-oriented variables in developing states, such as gross domestic product (GDP) per household with PC, significantly impact certification potential. On the other hand, all developed country learners' demographics, especially learners' age and levels of education, showed a high correlation with earning a certificate. Also, MOOC certificate earners were, as found in earlier studies on learner characterisations, such as Davis et al. (2013), of younger age groups, higher level of education, and enrolled mainly for professional development.

Similarly, Cagiltay, Cagiltay, and Celik (2020) statistically analysed course certification rates on edX, but with a larger dataset of 2.8 million learners of 122 MOOCs, the most extensive experiments in our survey in terms of the number of courses and learners. To analyse certificate attainment, the experiment used the courses' metadata (average chapters completed, total number of chapters, total forum messages, and learners' mean age). They found a positive relationship between the number of average chapters completed, mean age, the total number of forum messages and certification rates, in parallel with earlier findings by Hone and El Said (2016). In terms of course design, it was found that shorter and more interactive courses had higher certification rates. Discipline-wise, computer science and business courses were the most popular among learners. Most of the 3 million enrolled learners have degrees (bachelor's, master's, or both). These recent survey results are consistent with previous research on MOOC learners' higher education characteristics (Macleod et al., 2015; Bayeck, 2016; Christensen et al., 2013). These suggest that MOOCs are mainly designed to target professionals and that the fundamental motive of launching MOOCs, "democratising education", has been discarded (Reich and Ruipérez-Valiente, 2019; Cagiltay, Cagiltay and Celik, 2020). More interestingly, the general decrease in yearly certified learners, regardless of the increasing number of enrolled learners found by Cagiltay, Cagiltay and Celik (2020), is concerning and pending further research. Although the data analysed was sourced from only one platform, which may not be considered a general trend across MOOC platforms, the massive size of analysed learners and their heterogeneous background suggest examining the reasons for this phenomenon at a deeper level.

Samuelsen and Khalil (2018) statistically examined the correlation between the effort exerted over a specific time window and the likelihood of certificate attainment after achieving a final grade of at least 55%. The data were sourced via logs and a pre-course survey, including learner demographics (age, gender, level of education), learner's country of origin), video lessons, assignments, exams, and forum activities. This study examined the correlation between learners' effort and learning achievements and expectedly found that learners who exert more effort (specifically, more active days on the platform) have a higher probability of certification. However, learners with more than 100 active days have a lower probability of certificate attainment. Regarding learner demographics, female learners with higher degrees had a higher probability of obtaining a certificate, whereas age was negatively correlated with certification. As explained earlier, this study used the learner's number of active days to measure learner effort. However, this measure does not necessarily imply the effort exerted by the learner. Active time may indicate "active unattended sessions" rather than real active learning (Lee, 2018b).

Using two case studies, Littenberg-Tobias, Ruipérez-Valiente and Reich (2020) examined the impact of course price reduction by offering free certificate coupons on learners' certification behaviour. The first analysed the participation and certification rates in seven runs of four paid MOOCs versus two independent free-certificate-eligible MOOCs for comparison. There was a significant increase in the certification rate from 3% in the paid courses to over four times (13%) in the free-certificate-eligible MOOCs. There was no significant difference in both groups' demographics, apart from the country of origin, where the free certified learners were more likely to be from the United States. The second case study compared the behaviours and demographics of learners who purchased a computer science course against those who studied the same course after it was offered for free. Free learners were more likely to be women with a higher degree (mainly a PhD) and between 50-59 years old. The completion rates were 50% and 70% for free-certificate and paid-certificate courses, respectively, substantially higher than the completion rate in the standard post-paid courses, which stands at around 10% only. This study outlines learner behaviour differences in responding to MOOC price discounts.

Cobos and Jurado (2018) explored the impact of a three-dimensional perspective: learner's opinion (explicit attribute), interaction (implicit attribute), and context (contextual attribute) on obtaining a certificate (either free "honour" or paid "verified" at the learner's decision) at the end of the course, using two MOOCs of different disciplines (science and social science). While the descriptive analysis conducted in this study introduced a fundamental general insight into the learners' characteristics and behavioural patterns rather than a deeper statistically tested investigation of learners' behaviours, it outlines some general statistics of learners' certification behaviours in MOOC. From the total number of registered learners, the free (≈ 5%) and paid (≈ 3%) certification rate was higher in the social science course compared to the science course, where the free certification rate was around 4% and paid was around 1% only. This conforms with the findings of other studies, such as Wintermute, Cisel and Lindner (2021) and Littenberg-Tobias, Ruipérez-Valiente and Reich (2020). A synthesis of the certification statistics across all the surveyed studies is provided in Section 0. Other findings by Cobos and Jurado (2018), which include the course discipline-based difference in learner grades, time spent on assignments, time spent on videos, and learner degree, indicate that each course has its own characteristics. Hence, merging courses from different disciplines may not yield informative analyses and results.

Mullaney and Reich (2015) examined the influence of two different methods of course delivery (staggered versus all-at-one) on learner ontrackness, which was low in both paradigms. However, learners' persistence, completion, and participation were different. The sequential release showed learners stayed in cohorts, accessing the content through the course material over the first weeks in lockstep. Later, most learners started engaging in steps different from the most current (assigned) one. The all-at-one release paradigm was more asynchronous, in which almost all learners adapted their own individual pace through the course material. For certification, ontrackness modestly affected the certification positively, controlling for a learner’s number of active weeks, which was the strongest certification predictor in both courses but not in the same direction. The number of active weeks in the staggered course positively affected certification, whereas it negatively affected certification in the full release. According to the authors, a preliminary interpretation is that the all-in-one release allowed learners to reach the required certification score with less effort than the staggered course, where certificate-intended learners had to return to the course temporarily to keep attempting to reach the required certification grade score. Another interesting finding was that learners generally visit the first part of each course, the only portion of the course seen by a substantial number of learners. One recommended course redesign to entice learners is having an introductory substep each week summarising the main ideas and goals of the current week. The study concluded that releasing the course content in an all-at-one style is preferred for a more flexible learning experience. Course designers should consider this phenomenon while developing future runs or new MOOCs.

Goli, Chintagunta and Sriram (2019) studied the impact of paying for a MOOC certificate in advance, or within the first weeks of the course on learners’ engagement. This study, in particular, examines the effect of two temporal variables on learners’ engagement: (1) the effect of the certificate, which showed a boost in learners’ engagement until reaching the minimum required grade for earning a certificate, and (2) the effect of sunk-cost fallacy on paying but not certified learners, which is proved time transient and lasts only for a specific time after making the course fee payment. The analysis used 70-edX-course data and showed that paying learners engaged with the course and scored assessment marks higher than free learners to meet the certification condition (passing threshold). The above two variables increased paying learner engagement by about 10% compared to other non-paying peers. However, the difference between the two groups shrank as soon as the learner met the required certification threshold. The difference was more significant in the average final scores (73% for paying and 33% for non-paying), which interprets paying learners' commitment to the grade certification threshold. The variables used in the study included the learner's total time spent (minutes), average session duration (minutes), average number of sessions, forum activities (#posts/#visits), average grade, and graduation rate. Such findings can help platforms and course designers modify accordingly (e.g., redesigning course milestones or rescheduling fee payments, which are required by the first 24 days of the course) to maximise learners’ engagement with the course content and, at the same time, help monetise courses.

3.5.2. Machine Learning Models

3.5.2.1. Clickstream-based Models

Coleman, Seaton and Chuang (2015) used LDA to explore behavioural patterns via learners' click streams. In this work, learners’ interaction with the courseware was considered a “bag of interaction”, from which probabilistic use cases were formed for clustering learners based on their behaviours using LDA. Using the probability distribution associated with each case, an interpretable representation of access patterns for each user was formed, which helped predict the likelihood of certification. With little data, using click stream only, this model achieves promising results (0.81 ± 0.01 accuracy). Nevertheless, examining the effect of more factors, such as population size and course structure, on the resulting use cases may help improve the model performance.

Singhal (2023) predicted learners’ success using a HarvardXMITx-Course Dataset, which contains data for 641138. The proposed model is based on CNN, adopting eight possible input variables related to learners’ activities in MOOCs. The variables were tested as a preliminary step to identify their importance to the proposed predictive model; four inputs and two outputs were nominated to build the final version. The achieved performance of the proposed model ranged between 0.82 and 0.91 for predicting success and success level, respectively. The data adopted for building this predictive model include the number of daily activities, played videos, events, and the number of chapters opened.

3.5.2.2. Survey-based Models

Kostopoulos et al. (2021) recently built a multi-ML techniques-based certification predictive model and found that the weekly overall grades were the most predictor of certification. The 11-week dataset included some demographical variables (gender, employment status, current occupation) as well as some unexplored-before variables (professional experience in years, daily work hours, English language skills, digital proficiency skills, previous experience in MOOCs, mother tongue and average weekly available hours for study). The results show that boosting and ensemble were the top-performing algorithms, even without fine-tuning parameters (using the default values of parameters).

Rõõm, Luik and Lepp (2022) adopted a decision tree for predicting learners’ success – measured via computing the difference between learners’ intentions and their actual course performance – and certification in a computer programming MOOC. The factors (features) used to feed their model included learners' characteristics, engagement metrics and data collected with a voluntary questionnaire in a computer programming MOOC. The results showed that learners' prior education, prior experience with programming and online courses, use of the referred external materials and the motivation to obtain a certificate were the most influencing factors (representative features) for the predictive model to learn. The study concluded by suggesting complementing learning materials with links to external materials and developing a range of support mechanisms for learners to choose from, which may allow course providers to re-evaluate the resources used in the courses.

3.5.2.3. Discussion Forums-based Models

Jiang et al. (2014) employed learners' social interaction in Coursera's MOOC discussion forum to build a certification predictive model using the first week-only data. The features used included social network degree, which measures the local centrality of the learner interacting with his peers in the course discussion forum. This measure has already been calculated and explored in the authors' previous work (Jiang, Fitzhugh and Warschauer, 2014), whereas here, it was used for predicting certification using a logistic regression classifier. This study predicted not only certification but also the type of certificate earned, whether distinction or normal, with two prediction scenarios: distinction versus normal certificate and normal certificate versus no certificate.

Liu *et al.* (2022b) used the temporal cognitive topic model (TCTM) – an unsupervised learning method - to measure the influence of learners' cognitive engagement patterns (e.g. tentative or certain) and concerns (e.g. the topics in course content or logistics) on learner success in MOOCs. TCTM was specifically adopted to investigate learners' interaction while discussing different topics at different time points of the course using data from a Modern Etiquette course. The study found that certification acquisition and examination grades were tentatively discussed by low-achievement group (cluster), whereas high-achievement learners discussed more on-task topics. Moreover, a moderation analysis revealed that the moderating effect of discussion guidance, especially for instructor-led guidance, between learning achievements and salient cognitive topics was significant.

3.5.2.4. Multi-sources-based Models

This section reviews studies that employ machine learning for predicting certification in MOOCs, either using supervised or unsupervised learning. For instance, Ruipérez-Valiente et al. (2017) used a combination of raw and computed features to build the predictive models, which include the grade achieved in the assignments (problem progress), the percentage of the video watched (video progress), the total time spent on assignments (problem time), the total time spent on watching videos (video time), the total time spent on the whole course (total time), the total number of course visits (sessions), the number of events produced by the learner (events), the total number of days that learners logged (logs) and the time invested in each day of the course (constancy). These studies provide temporal (weekly-based) performance results and suggest focusing on the learners' first four-week course activities only to build an effective predictive model. It also found that features, in terms of importance, play a massively different important role over the weeks; hence, early learner warning models should tune the weight of the variables accordingly during the different weeks of the course.

Lee (2018b) used weekly-based features gathered from different sources, including videos, discussion boards, wikis, weekly assignments, quizzes, and midterm and final exams, to examine the effect of the uninterrupted time-on-task variable on certification. The present study attempts to address the claim that the previous predive model's file logs could not recognise the off-task time during an active MOOC learning session (i.e., while the web browser window remains open but the learner is not present). To address this issue, sessions that are longer than pre-determined thresholds (10, 30, and 60 minutes) are excluded from the analysis. The findings unsurprisingly showed a positive correlation between learners' weekly learning activities and learning sessions and earning a course certificate. Interestingly, the likelihood of obtaining a certificate increases when more learning activities are performed in fewer sessions. This suggests that learning activities alone may not be sufficient to examine how learners self-regulate their learning, and activity grouping would help form more meaningful learning experiences.

The same author later used the same course data (Lee, 2019) to cluster learners to identify different groups of similar-characteristics learners using their weekly homework and quizzes. The clustering techniques include hierarchical clustering algorithms (HCA) and self-organising maps (SOM) and revealed several learning patterns based on various activities (e.g., number of attempted quizzes [problem-solving activity], weekly problem completion percentage, and certificate attainment). Next, a random cluster with an almost equal number of certified and non-paying learners was further analysed using logistic regression. As expected, certified learners attempted more problems compared to their peers. Holding all other predictor variables constant at their mean values, the log odds for certification increase when the learner solves one more weekly quiz or homework. The same positive correlation with certification was also observed with other variables (e.g., problem completion percentage on the due date). However, the interesting finding in this study is how the weekly problem completion percentage negatively correlated with certification in the first two weeks of the course. According to the authors, one possible interpretation is that these learners exerted more effort when the course was easier. In addition, the weekly problem completion percentage started dropping from week four to the end of the course, which may also represent their despair in certificate attainment.

Qiu et al. (2016) examined the correlation between learners' demographics, forum interactions, behavioural patterns, and certification using data from the only Chinese MOOC platform in this survey XuetangX. One interesting finding is that learners who ask questions are more likely to complete the course and obtain a certificate than those who answer questions. This triggers an intention towards the scalability of MOOC forums and whether they have been effectively used either by learners to learn beyond the official content or by researchers for modelling certification. Our survey shows very little natural language processing (NLP)-based analysis of the content of the forums, where most of the studies only used numerical variables (e.g., the number of posts and replies) for prediction. Another finding of this study is the significant relation between learner level of education and course discipline. This shows that non-degree learners are more likely to enrol in none-science courses, while graduate learners are likelier to enrol in science courses. Gender-wise, female learners enrolled and obtained certificates at a higher rate in economics, history, and sports courses compared to computer science and engineering, which were dominated more by male learners. Regarding the effort exerted by learners, graduates did fewer activities and spent less effective learning time but achieved a higher certification rate. This indicates that a learner's knowledge is essential in increasing the likelihood of certificate attainment.

Xu and Yang (2016) employed learners' click stream activities (video watching) to cluster learners based on their motivation and then predict learners' grades and certification. Using 10 edX courses, the model was tested on only support vector machine (SVM) but with four different kernels and reported performance's general accuracy only.

Tian et al. (2017) clustered learners into three grade-based groups and analysed learners' behaviours statistically using learner activities: events, time spent in days, video watching, steps, and forum activities. K-means was used to find separability between the different categories. A further relationship analysis was conducted using learner categorisation based on learner access (registered: learners who have never accessed the courseware; active: access more than half of the course; general: represents the remaining learners). The three groups' percentages were roughly 37%, 57%, and 6%, respectively. Later, a certificate-obtaining predictive model was developed using LDA, LR, and SVM classifiers.

Cobos and Olmos (2018) observed the traditional challenge of MOOCs, that despite the massive number of enrolled learners, courses still face many dropouts and low end-of-course certificate-earning rates. However, there are several reasons behind this phenomenon that MOOC providers should consider:

- Many learners join MOOCs as 'curious learners' having no initial intention to complete the course or earn a certificate.
- Learners do not spare enough time for self-learning activities.
- Loss of interest in the course content due to various factors (e.g., difficulty, avoiding exams, or a stressful environment, specifically for certification).
- The financial unwillingness to pay for the certificate, even if the certification requirements are met (Cobos and Olmos, 2018).

This study used edX data from 15 runs of seven MOOCs to model certification and develop a Model Analyser System Plus for edX MOOCs (edX-MAS+) to classify learners as dropouts or certificate earners. They found that the Bayesian Generalized Linear Model and Stochastic Gradient Boosting performed the best, followed by simple artificial neural network (ANN) and random forest (RF). This suggests that deep models do not necessarily perform better for MOOC predictive modelling, considering the data size or the data linearity, where learners tend to have lower interaction towards the end of the course.

Gitinabard et al. (2018) combined learner click stream and discussion forum data from two runs (offerings) of the same course to build a certification predictive model. The conducted features importance analysis indicated that learners' total number of attempts, video views, votes (likes and dislikes of other comments), and posts were the most important features for predicting learner likelihood of certificate attainment. Thus, behavioural features were better predictors of certification than social behaviours, which might be because, typically, only a small group of learners participated in the course discussion forum. This study followed a sound methodological approach by training the predictive model on the course's first offering and testing the model performance using the latest offering available, practically simulating the real-world scenario in which MOOCs are offered in consecutive annual runs. However, the data were balanced (the none certificate earners class was down-sampled to match certificate earners) before training the model, which may challenge the model's reliability. Certification predictive models should deal with MOOCs' highly imbalanced datasets and provide real-life compatible solutions.

Fotso *et al.* (2022) used various predictive algorithms, recurrent neural networks (RNNs), long short-term memory (LSTM) and gated recurrent unit (GRU) to predict outcomes and learners' interactions in MOOCs. The data adopted in their experiment was obtained from UNESCO's International Institute for Capacity Building in Africa, which designs MOOCs for teacher training. The variables used to build the models include social, geographical, and learning behaviours. This experiment is among the very few that fine-tuned the model parameters using L2 regularisation to improve the accuracy of the predictive model. The findings revealed that RNN was the best-performing predictor compared to the other deep learning architectures. Additionally, a correlation between video viewing, quiz answering, and the level of learner participation was observed. Also, the video or quiz length correlated with the viewing behaviour, where shorter videos and quizzes score a higher number of viewing (interaction). Such findings indicate the need for extensive analysis of designing futuristic courses regarding videos and quizzes.

3.6. Synthesis of The Surveyed Works

This section introduces a high-level outline and synthesis of the surveyed MOOC certification predictive models. We profile the data sources (platforms, course providers, number of courses/runs, number of learners, course delivery interval, and type of data used), adopted methodologies (algorithms/models, performance metrics), and model outputs (model prediction earliness, certification type, (i.e., free or paywalled) for better categorisation and synthesised analysis of the works surveyed.

3.6.1. Data Sources

3.6.1.1. Platforms

The data sources for MOOCs prediction modelling have not been within the attention of several previous reviews on MOOCs research (Gardner and Brooks, 2018b). Exploring the sources (platforms) utilised for building MOOCs' predictive models helps us understand how platform data plays a role in the current certification models. Our analysis shows that edX-based predictive models were overwhelming, where 60% (n = 15) of the surveyed works used data from edX, as shown in

Figure 3.5.

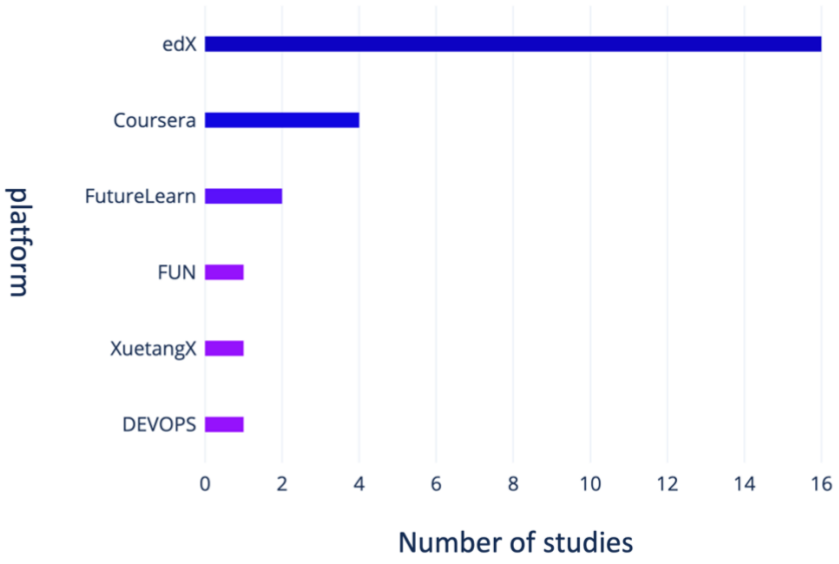
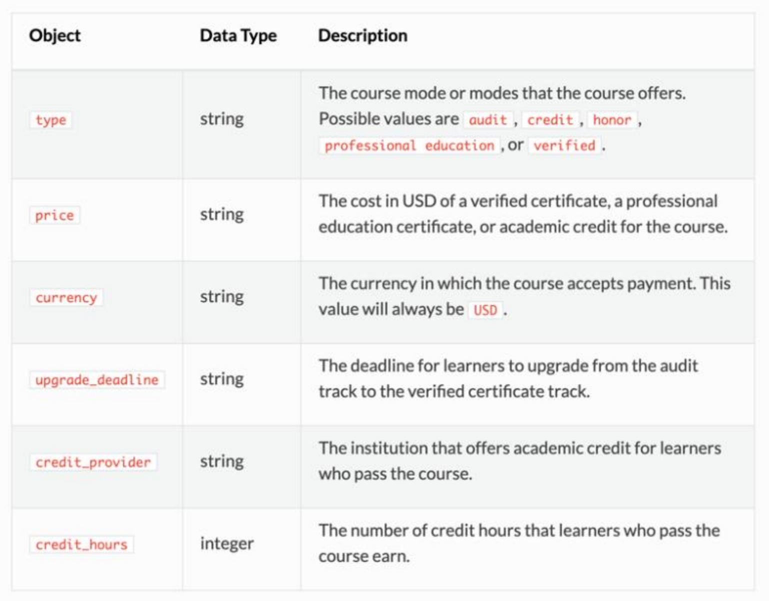


Figure 3.5. Distribution of the platforms from which the surveyed models obtained data.

There is no specific explanation for this trend; nevertheless, we estimate that the ease of obtaining an edX API[43] and access to some course financial details has played a role in this dominance. In addition, the data edX provides access to various course financial variables, as shown in Figure 3.6. Thus, this may have increased researchers' intention in modelling certification using edX data.

[43] https://course-catalog-api-guide.readthedocs.io/en/latest/index.html



Object	Data Type	Description
`type`	string	The course mode or modes that the course offers. Possible values are `audit` , `credit` , `honor` , `professional education` , or `verified` .
`price`	string	The cost in USD of a verified certificate, a professional education certificate, or academic credit for the course.
`currency`	string	The currency in which the course accepts payment. This value will always be `USD` .
`upgrade_deadline`	string	The deadline for learners to upgrade from the audit track to the verified certificate track.
`credit_provider`	string	The institution that offers academic credit for learners who pass the course.
`credit_hours`	integer	The number of credit hours that learners who pass the course earn.

Figure 3.6. A snapshot of the financial variables provided by the edX API.

Concerning other giant platforms, Coursera, in contrast, has deprecated its API[44] for unknown reasons and declared that the intellectual property of the course content and learner data are owned only by the partner institution (course provider)[45]. This is also the case with the most prominent European MOOC platform FutureLearn[46] where learners' data are also under the management of the course providers. As illustrated in

Figure 3.5, other providers were less represented (e.g., the French FUN, the Chinese XuetangX, and the Greek DEVOPS). Non-English-speaking learners form a considerable portion of MOOC participants and typically have different behaviour patterns, which require further analysis. These platforms may have sourced more literature on certification modelling than we surveyed in this work; nevertheless, they were excluded earlier in Section 3.4.1 for not being authored in English.

3.6.1.2. Publication Years and Numbers of Learners/Courses/Runs

[44] https://build.coursera.org/app-platform/catalog/old.html
[45] https://www.coursera.org/about/terms
[46] https://www.futurelearn.com/info/terms/privacy-policy

The publication years ranged from 2014 to 2021, as shown in Figure 3.7 whereas the number of courses analysed in our surveyed works ranged from one to 140. The size of the nodes in the figure below denotes the number of learners in each work, ranging from only 961 learners in Kostopoulos et al. (2021) to 2.8 learners in Cagiltay, Cagiltay and Celik (2020). Although the number of learners showed no association with the publication years, we observed a positive association between the number of courses and the publication years, as shown in Figure 3.7.

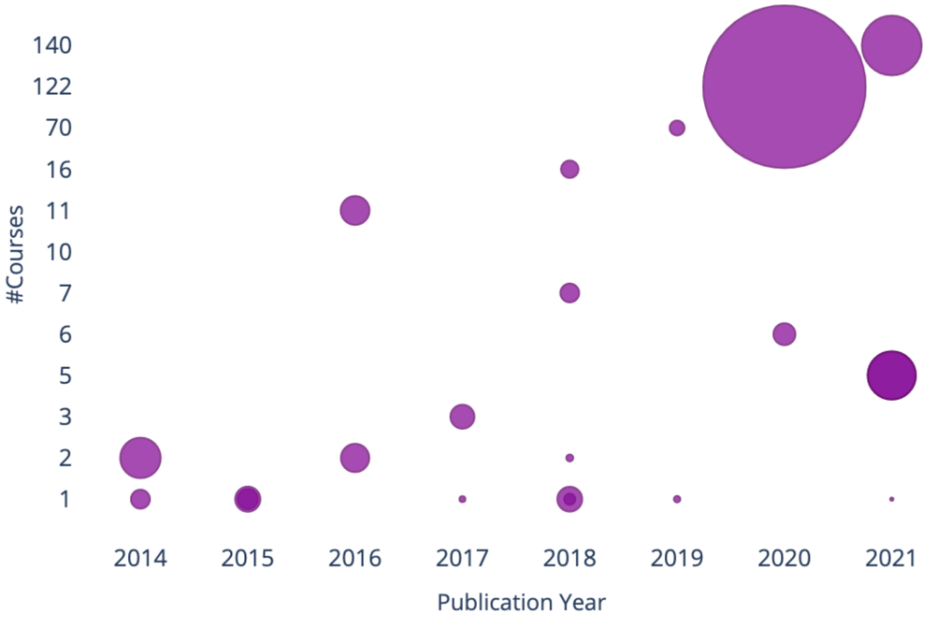


Figure 3.7. The number of courses in each surveyed study distributed by publication year. Node colour darkness denotes the number of studies at that value (number of courses and publication year).

Figure 3.8 illustrates the years covered by our surveyed works (the course delivery intervals). Eight works used data from single-year-delivered courses, whereas around 11 studies (60% of the course delivery interval stated models) used temporal data ranging from 2 to 5 years.

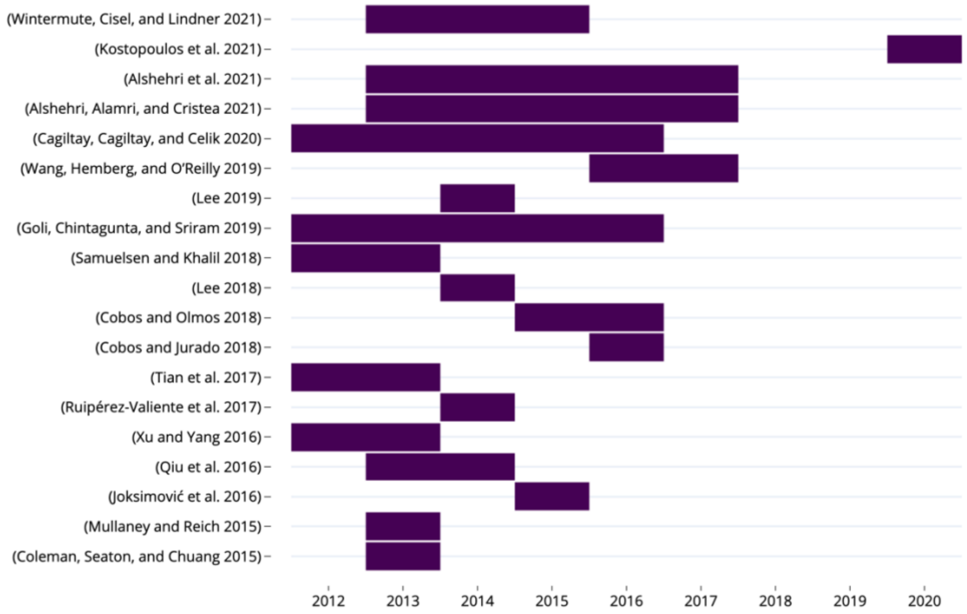


Figure 3.8. The surveyed models distributed by the course delivery intervals. Note: studies with undefined delivery intervals are excluded from this figure.

Using different types and data sources typically helps yield better performance of the predictive models; nevertheless, we observed an expected inverse relationship between the amount of data used (i.e., the number of features employed) and the number of learners involved in the experiment. This is expected when researchers non-optionally lose more data due to dealing with missing values while adding more input features to their predictive models. For instance, Kostopoulos et al. (2021) used unique (non-previously explored) features to predict certification. However, the number of included learners was relatively low, with the lowest (n = 961) among our surveyed works. Another example is (Joksimović et al., 2016), where the authors dropped the survey-obtained data because it was not completed by most learners and sufficed with the discussion forum data (comments and replies). This allowed the authors to apply their model to a relatively large number of learners (almost 85,000).

The data types employed by the surveyed works for predicting certification range from the most common data type of clickstream used in over 70% of the surveyed studies to the lowest-used source of pre-course surveys, as demonstrated in

Figure 3.9. This is not surprising because clickstream data:

- Are rich in information about learners and granular at various levels.

- Do not require more extensive human and computational pre-processing than other types, such as learner textual data (comments and replies).
- Usually contain the data of all enrolled learners.
- Are better predictors of MOOC learners' behaviours compared to other data sources (e.g., clickstream, forum discussions, and assignments) (Gardner and Brooks, 2018a).

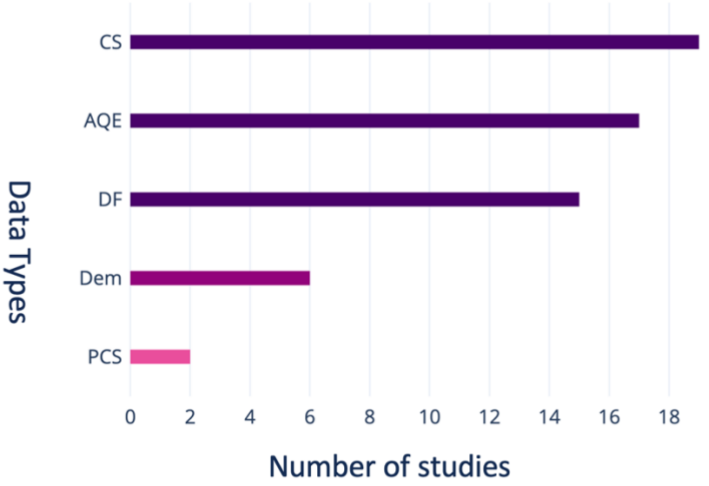


Figure 3.9. Data types used in the surveyed studies.

3.6.2. Adopted Methodologies

3.6.2.1. Data Pre-processing and Features Engineering

One of the core steps of feature engineering is feature selection, which was a crucial step in several surveyed studies not only for identifying the most representative features for prediction but also for improving the model performance. In contrast to the general tendency towards having more learners' data, which was briefly discussed in section 3.6.1.2 and more detail in this review's limitation section, the features used for building a predictive model should be more effective than many. However, not all the studies surveyed have explained the feature selection step. This might be due to the relatively limited number of already available features; hence, training the model on the whole dataset was the option. For example, Ruipérez-Valiente et al. (2017) selected 11 variables to build the predictive model without explaining whether a feature selection had been conducted.

Another reason for skipping feature selection is following statistical analysis rather than a machine learning approach. These studies tend to conduct descriptive statistics and correlations to measure the association between each variable (feature) and certification. For instance, Cagiltay, Cagiltay and Celik (2020) used a massive dataset of 2.8 million learners' activities (average chapters completed, total number of chapters, total forum messages) and demographics (learners’ mean age) to measure the association between these variables and certificate attainment. Besides other statistical analysis-based studies, this study typically aims to evaluate the correlation between the available variables and certification without further ML; hence, a feature selection may not be applicable in this scenario. This is the case with other statistical analysis-based studies surveyed, especially those that used very few sources or data types, such as Lee (2018b), who examined the effect of uninterrupted time on tasks on certification.

The model's initial results, being promising due to the nature of the dataset itself, might be another reason for skipping the feature selection step. For example, Kostopoulos et al. (2021) reported high-performance metrics of their model with an area under the curve (AUC) over 0.90 across all the algorithms employed and using only the course first-week data. Therefore, feature selection may be intentionally dropped due to the unnecessary from the author(s) perspective. However, this study later reported the importance of each feature to model performance using Shapley Additive exPlanations (SHAP) via plotting the SHAP value for each feature. This technique visually illustrates how much each feature has influenced the classifier’s decision. Another feature selection-based study is Gitinabard et al. (2018), where features with an importance measure of one or more were selected.

Few studies have dedicated a separate subsection to feature selection. For instance, Cobos and Olmos (2018) developed a Model Analyser System for edX MOOCs (edX-MAS+), which automatically extracts, cleans, selects, and pre-processes course data (features) and makes them ready for feeding to the model. The tool automatically compiles several functionalities for processing the MOOC data to obtain the representative input variables before feeding them to the ML algorithms adopted for the prediction task. Thus, feature selection is part of the edX-MAS+ workflow. However, no further details on the method of computing feature importance were explained.

3.6.2.2. Hyperparameter Fine-tuning

Although configuring predictive algorithm parameters is essential for improving model predictability, most surveyed works skipped this step or simply let the model assign the default parameters. Having the parameters tuned generally helps the model achieve better-forecast results and find and diagnose common modelling issues such as bias, underfitting, and over-fitting. Due to each experiment's different nature, this step improves the accuracy of the forecasted results. Some studies mentioned that tuning algorithm parameters were part of building their models without further details on what parameters and how they were tuned (Cobos and Olmos, 2018; Gitinabard et al., 2018).

Kostopoulos et al. (2021) tuned some parameters in their eight adopted classifiers, mainly the learning rate and the minimum number of samples to split node, but most of the classifiers were used with the default values of the parameters. In particular, ensemble methods were used “without any special parameter configuration”, as stated by the authors. (Qiu et al., 2016) similarly reported tuning the parameters of one of the classifiers used (Latent Dynamic Factor Graph [LadFG]), which was the best performing compared to other algorithms (LR, SVM, factorisation machine [FM]). Nevertheless, the default-parameters-based performance of the baseline model was not reported; hence, we cannot determine the extent to which the parameter tunings helped improve the results. Tian et al. (2017) also reported tuning a few parameters, mainly related to SVM, which significantly improved based on various model deployments.

The limited discussion and reporting of tuning parameters in the surveyed studies might be justified by the uncomplexity of predictive models and the limited time and computational resources MOOC models generally consume. However, we expect the surveyed works to be more specific on tuning parameters if more costly data (e.g., course video content or learner-generated textual data) were analysed, which was not done among the works surveyed in this review.

3.6.2.3. Models/Algorithms

Figure 3.10 demonstrates the modelling algorithms employed across the surveyed works. The main approaches followed involve statistical analysis (including descriptive statistics, correlation, and regression analysis), supervised machine learning (with only one study adopting deep learning), and unsupervised machine earning (clustering). Additionally, they include some other limitedly used algorithms/models: Bayesian models such as Bayesian Generalised Linear Model (BGLM), matrix decomposition such as Latent Dirichlet Allocation (LDA) and FM, discriminant analysis such as LDA2, and dimensionality reduction such as Self-organising Map (SOM).

Our study shows that regression was the most common technique, followed by ensemble-and tree-based algorithms at an almost equal level of adoption. The little scope of DL-based models may be interpreted by the general linearity of MOOC data because (1) the majority of platforms log learners' data on a weekly basis because MOOCs are minimally structured by design into consecutive weekly-based learning units (Yeomans, Reich and Acm, 2017; Wang, Hemberg and O'Reilly, 2019) and (2) there is a general decline of learner activities towards the end of the course. In parallel with this, recent experiments have found that complex deep models to classify MOOC learners may not necessarily boost prediction performance or overperform conventional ML algorithms (Aljohani and Cristea, 2021; Sebbaq, 2022). Regarding the below figure, multiple-algorithm works are counted multiple times. See Table 3.4 for the complete list of metric abbreviations and acronyms.

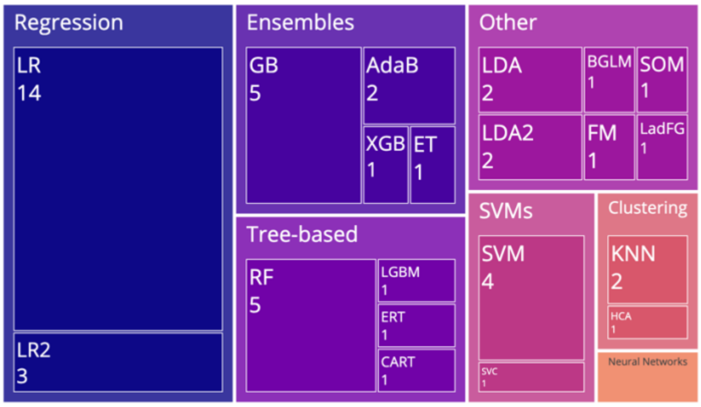


Figure 3.10. The modelling algorithms employed in the surveyed works clustered by model type.

The performance metrics reported in the surveyed works vary, as demonstrated in Figure 3.11, based on the algorithm/model used. We can see that cross-class metrics such as accuracy (Acc.), F1, and AUC have been mainly reported. This diversity of metrics is typical, considering that different metrics evaluate different aspects of performance quality, which vary based on the data used, the methodology followed and the research objectives. For example, eight studies (over one-third of the surveyed studies) reported accuracy solely or with other metrics. While accuracy, which works well when the number of samples belonging to each class is equal, may be informative in many classification tasks, MOOC certification datasets are usually highly imbalanced; hence, such a threshold-dependent metric may be misleading. Other metrics that evaluate performance over all possible thresholds, such as AUC, should be adopted more by MOOC certification predictive models. However, reporting appropriate performance metrics often depends on the outcome being measured and the objective of the prediction task. For instance, recall

(Rec.) may be the appropriate metric when the goal is a simple intervention (like reminding the learner), whereas precision (Prec.) might be the best metric for resource-intensive support to predict certification (Gardner and Brooks, 2018b).

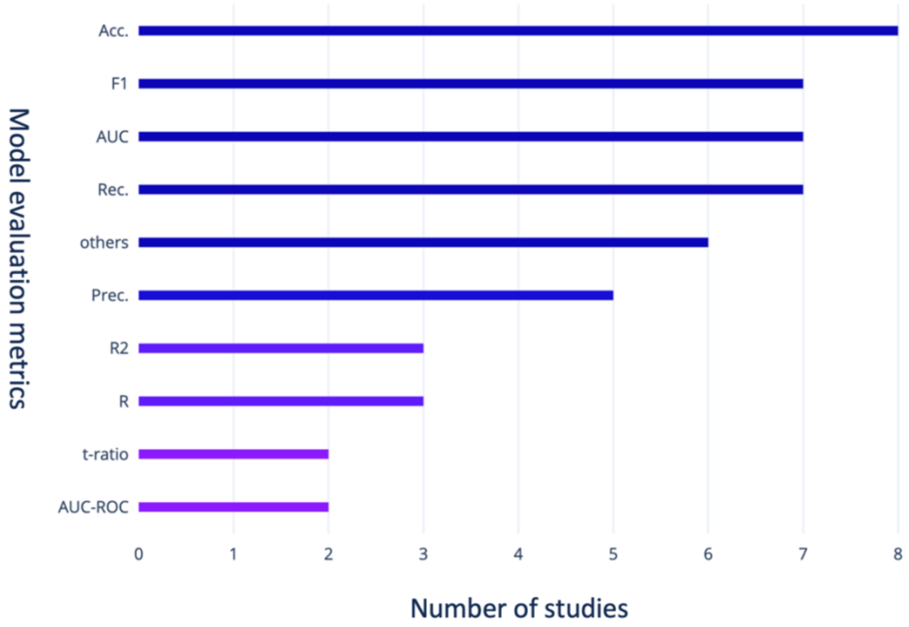


Figure 3.11. The model evaluation metrics reported for the surveyed works. Note: multiple-metric works are counted multiple times.

3.6.3. Model Outputs

3.6.3.1. Types of Certifications

The surveyed studies reported a shallow certification rate, from less than 1% to only 3% of the registered learners for paid courses (Littenberg-Tobias, Ruipérez-Valiente and Reich, 2020; Cobos and Jurado, 2018; Alshehri *et al.*, 2021) purchase a certificate, and around 4.5% to 13% of the enrolled learners for free courses (Ruipérez-Valiente et al., 2017; Littenberg-Tobias, Ruipérez-Valiente and Reich, 2020; Cobos and Jurado, 2018; Wintermute, Cisel and Lindner, 2021). These statistics are lower than the MOOC completion rate (broadly around 10%) due to the additional requirements of certificate attainment for freemium courses (reaching a pre-defined threshold, of course, passing grade) and premium course certification. However, this low certification rate in MOOCs should not be interpreted as either the content being difficult to follow or the learners being unwilling to invest in learning. On the contrary, many learners are driven by the intellectual stimulation offered by MOOCs and want to use these courses for complementary lifelong learning, learning new skills, or simply exploring what learning is like outside of the confines of an institution (Kizilcec, Piech and Schneider, 2013).

Regarding the course passing grade, a requirement for obtaining a certificate in most MOOC platforms, just over 30% of learners achieve the minimum passing grade (Goli, Chintagunta and Sriram, 2019). MOOC learners not only leak out at later stages of the course but also massively, even before the course commences. Learners enrolled in their course after registration are surprisingly less than half of the learners (only 46% to 60% of the total number of registered learners) (Alshehri, Alamri and Cristea, 2021; Cohen et al., 2019; Cagiltay, Cagiltay and Celik, 2020; Reich and Ruipérez-Valiente, 2019). Based on the above statistics, Figure 3.12 demonstrates that learners leak out through the course from a certification perspective.

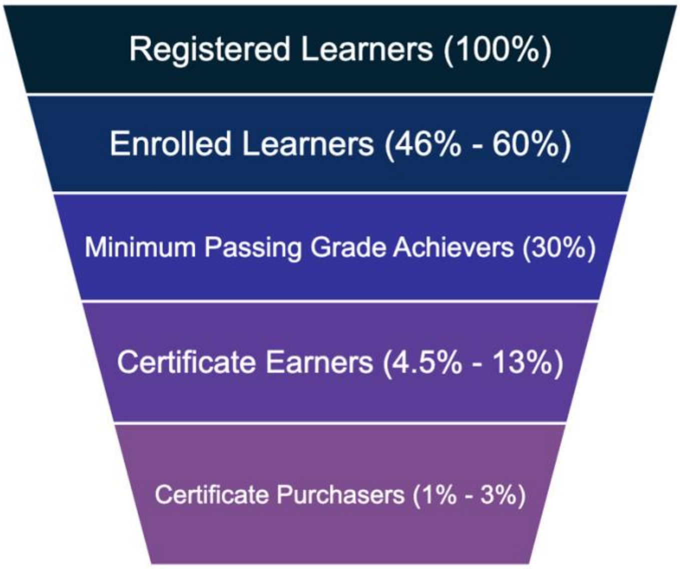

Registered Learners (100%)
Enrolled Learners (46% - 60%)
Minimum Passing Grade Achievers (30%)
Certificate Earners (4.5% - 13%)
Certificate Purchasers (1% - 3%)

Figure 3.12. MOOCs certification leaks out towards the end of the course. Percentages are rounded and represent the total number of registered learners.

Paid certification concerns modelling certificate purchasers' behaviours and is a unique research area. This trend coincides with recent platforms' introduction of more paywalled courses, such as university-affiliated online degrees and corporate training. The surveyed studies of this concept attempt to model learners' financial decisions by determining the most representative features for modelling course certificate purchase behaviour and building predictive models.

3.6.3.2. Early Prediction Models

Early prediction of certification is vital to detect learners' behaviours and provide them with timely support and effective intervention (Kostopoulos et al., 2021). In line with this concept, 40% (n = 10) of our surveyed studies addressed the need for early prediction of certification in MOOCs, but from a different perspective. For instance, Cobos and Olmos (2018) used MOOC temporal data of learners' daily and weekly activities to predict learner certificate attainment. Cagiltay, Cagiltay and Celik (2020) similarly built their experiment on a dataset that contains variable measures at three-time points of the course rather than the widely adopted weekly-based modelling: (1) at the beginning of interacting with the course content, (2) at the half-point of the course and (3) after finishing the course. This time-scale data dimension and the massive number of analysed learners of 2.8 million helped build a rich and fine-grained representation dataset. However, this dataset was recorded at a meta-level, which included measurements at the course level rather than of learners. Consistent with previous studies that investigated the importance of early prediction (Hone and El Said, 2016) (Cagiltay, Cagiltay and Celik, 2020), the likelihood of certification increases when a learner passes the midpoint of a course.

Early prediction is essential for informing instructors of the estimated learners’ behaviour to provide them with relevant advice for better progress with the course (Cobos and Olmos, 2018). On this basis, Ruipérez-Valiente et al. (2017) suggested that early prediction should be associated with implementing an adaptive educational system for adapting the course content based on the learner's predicted behaviour. This study, in line with other early predictive models, suggests focusing on the first weeks of the course, specifically the first four weeks, to build an effective predictive model. Kostopoulos et al. (2021) used an 11-week MOOC for early prediction and found that weekly grades were the most predictive features. It shows that boosting and ensemble were the top-performing algorithms, even without any parameter fine-tuning performed (using the default parameter values).

3.6.3.3. Computational and Time Cost of Modelling

Only one study reported the performance of the employed algorithms from the time-costing perspective (Cobos and Olmos, 2018), which is, at the same time, the only study that used deep learning, along with other traditional ML algorithms, for modelling. The study noted that neural networks, extreme gradient boosting, and random forest were the highest time-consuming algorithms for training, whereas NB took the most time for prediction (testing). As stated earlier, MOOC prediction models tend to use shallow learning algorithms. Therefore, the computational and time cost may not be a challenging variable compared to more complex deep learning tasks such as pure NLP or computer vision.

3.6.3.4. Model Error Analysis

Error analysis isolates and diagnoses the model's classification errors into meaningful outputs and highlights the most frequent errors, along with the characteristics (input features and observations) associated with the model misclassification. Similar to feature selection and hyperparameter configuration, error analysis was rarely reported; only two surveyed studies have analysed their model's erroneous classifications and reached the same conclusion. Moreover, Qiu et al. (2016) and Tian et al. (2017) found that the similarity between both groups (certified and noncertified learners) plays an essential role in model classification errors. They found that a significant proportion of the learners were actively engaged with the course content and, in the forum, behaved like certificate earners but did not obtain certificates. These learners were misclassified into the certified group. The opposite has also happened, where some inactive certified learners were misclassified. Additionally, the minimum score for certificate eligibility was found to be one of the challenging variables of the model, where many learners with scores hovering around the minimum score were misclassified.

Table 3.2. Outline of previous studies on MOOC certification prediction (n/a: missing, EM: Early Model, CT: Certification Type, other abbreviations are explained in the following tables)

Ref.	**Platform**	**Provider**	**Delivery Interval**	**#Courses: #Runs**	**#Learners**	**Data Type**	**Models/s Algorithms**	**Performance Metrics**	**EM**	**CT**
(Cobos and Olmos, 2018)	edX	The University Autónoma of Madrid	2015 – 2016	7:15	37,420	CS; DF; AQE	LR, GB; SVM; K-NN, RF; NN; NB; BGLM;	AUC; ROC	Y	n/a
(Arslan, Bagchi and Ryu, 2015)	edX	Harvard & MIT	n/a	n/a	358,433	CS; Dem.	LR2	R; β; SE; *t*-ratio	N	n/a
(Cagiltay, Cagiltay and Celik, 2020)	edX	MIT	2012 – 2016	122:n/a	2.8 million	CS; DF	LR2	R; β; SE; *t*-ratio	Y	n/a
(Ruipérez-Valiente *et al.*, 2017)	edX	Universidad Autónoma de Madrid (Spain, UAM)	2014	1:1	3,530	CS; AQE	RF; GB; kNN; LR	AUC; F1; Acc.	Y	Paid
(Lee, 2018b)	edX	MIT	2014	1:1	12,981	CS; DF; AQE	LR	AUC; Rec,	N	Free
(Goli, Chintagunta and Sriram, 2019)	edX	n/a	2012 – 2016	70:n/a	23,674	CS; DF; AQE	LR2	M, P	N	Paid
(Qiu *et al.*, 2016)	XuetangX	n/a	2013 – 2014	11:n/a	88,112	CS; DF; Dem.	LadFG; LR; SVM; FM	AUC; Prec.; Rec.; F1	N	n/a
(Xu and Yang, 2016)	edX	Harvard & MIT	2012 – 2013	10:n/a	n/a	CS; DF	SVM	Acc.	N	n/a
(Alshehri, Alamri and Cristea, 2021)	FutureLearn	Warwick University	2013 – 2017	5:23	245,255	CS; AQE	RF, ET, LR, SVC	Acc.; Rec.	Y	Paid
(Tian *et al.*, 2017)	edX	Harvard & MIT	2012 – 2013	11:n/a	n/a	CS; DF; AQE	LDA; LR; SVM	Acc.; Prec.; Rec.; F1	N	n/a

(Samuelsen and Khalil, 2018)	edX	Harvard & MIT	2012 – 2013	16:n/a	32,621	CS; DF; AQE	PC; LR2	β; SE; CI	N	Free
(Littenberg-Tobias, Ruipérez-Valiente and Reich, 2020)	edX	Harvard & MIT	3 years	6/12	50,927	CS; DF; AQE	LR2	R2, F2, SE	N	Paid
(Alshehri *et al.*, 2021)	FutureLearn	Warwick University	2013 – 2017	5/23	245,255	CS; Dem.	RF, GB, AdaB, XGB	Acc.; Prec.; Rec.; F1	Y	Paid
(Wintermute, Cisel and Lindner, 2021)	FUN	Universite´ Nume´ rique	2013 – 2015	140:n/a	378,000	CS	LR2	β	N	Free
(Cobos and Jurado, 2018)	edX	The University Autónoma of Madrid	2016	2:2	5,011	CS, Dem; DF; AQE; PCS	DS	DS	N	Free + paid
(Yeomans, Reich and Acm, 2017)	edX	Harvard & MIT	n/a	3:3	60,778	PCS	DS; LDA2; LR2	DS, β; CI; AUC; R2; Z; SE	Y	n/a
(Mullaney and Reich, 2015)	edX	MIT	2013	1:2	66,774	CS; Dem.	LR2	Log; R2; SE	N	n/a
(Wang, Hemberg and O’Reilly, 2019)	edX	MIT	2016 – 2017	2:6	n/a	CS; AQE	DS	DS	N	Paid
(Lee, 2019)	edX	MIT	2014	1:1	4,337	n/a	HCA; SOM; LR	R; SE	Y	n/a
(Kostopoulos *et al.*, 2021)	DEVOPS	University Of Thessaly, Greece.	2020	1:1	961	CS; Dem.; AQE; DF	AdaB; GB; CART; ET; LDA; LGBM; LR; RF	Acc.; AUC; Prec.; Rec.; F1	Y	Free
(Jiang, Fitzhugh and Warschauer, 2014)	Coursera	University of California	n/a	2:1	173,000	DF	Permutation Test (corr)	R	N	n/a

(Jiang *et al.*, 2014)	Coursera	University of California	n/a	1:1	37,933	DF	LR	Acc.; ROC; Prec.; Rec.; F1	Y	Free
(Coleman, Seaton and Chuang, 2015)	edX	MIT	2013	1:1	43,758	CS	LDA2	Acc.; Rec.	Y	Free
(Joksimović *et al.*, 2016)	Coursera	The University of Edinburgh & ORT University Uruguay	2015	2:2	84,786	DF	LR2	P; SE	Y	n/a
(Gitinabard *et al.*, 2018)	Coursera; edX	Columbia University	2013 - 2015	1:2	65,203	CS; DF	LR	AUC; F1	Y	n/a

Table 3.3. List of data type abbreviations and acronyms

Abbreviation / Acronym	Description
AQE	Assignments/Quizzes/Exams
CS	Clickstream
Dem	Demographics
DF	Discussion Forum
PCS	Pre-course Survey

Table 3.4. List of approaches, abbreviations, and acronyms

Abbreviation / Acronym	Description
AdaB	Adaptive Boosting
BGLM	Bayesian generalised linear model
CART	Classification and Regression Tree
ET	Extremely Randomised Trees
FM	Factorisation Machine
GB	Gradient Boosting
HCA	hierarchical clustering algorithms
K-NN	K-Nearest Neighbour
LadFG	Latent Dynamic Factor Graph
LDA	linear discriminant analysis
LDA2	Latent Dirichlet Allocation
LGBM	Light Gradient Boosted Machine
LR	Logistic Regression
LR2	Linear Regression
NB	Naïve Bayes
NN	Neural Network
PC	Pearson's Correlation
RF	Random Forest
SOM	Self-organising map

SVC	Support Vector Classifier
XGB	XGBoostnig

Table 3.5. List of metric abbreviations and acronyms

Abbreviation / Acronym	Description
Acc.	Accuracy
AUC	Area Under Curve
CI	Confidence Interval
DS	Descriptive Statistics
F1	F1-score
F2	f-test (f statistic)
M	Population mean
P	p-value
Prec.	Precision
R	Correlation Coefficient
R2	Coefficient of Determination
Rec.	Recall
SD	Standard Deviation
SE	Standard Error
β	probability

3.7. Limitations

This section highlights the current MOOCs certification predictive models in terms of limitations and potential improvements. We reviewed the surveyed works from various aspects and reported methodological concerns such as generalisability, extensive experimental filtration, model replicability, and explainability. We also discussed opportunities for future research and improvement.

3.7.1. Model Generalisability

The “need for more data for a higher level of model generalisation and further validating the achieved results” was the most stressed call by the surveyed studies. As demonstrated earlier in Figure 3.7, a considerable number of studies (n = 14/25) have based their findings on a few courses (from one to three courses only); thus, it is challenging to consider the findings of these models generalisable. Furthermore, learners’ behaviours and certification rates differ based on the subject and discipline of the MOOC (Cobos and Jurado, 2018). Therefore, building the model on a diverse dataset would help increase the findings’ generalisability and reliability of results. Nevertheless, one model’s performance cannot be compared to others due to the different characteristics of each single MOOC. For instance, learner demographics and activities tend to have different statistics based on various factors, such as the MOOC discipline; thus, a model applied to a specific dataset may not be suitable for static application to another dataset (Samuelsen and Khalil, 2018). This is truer when further variances are introduced (e.g., when two datasets are of different platforms), where the features are typically not unique.

Thus, while training models on a domain-diverse dataset can help reach more reliably generalisable results, the results themselves should be interpreted with caution due to any potential biases, where the learners’ activities’ dataset used within an experiment may be filtered uniquely or potentially represent that certain MOOC provides only; hence, results may not be suitable for generalizability (Whitehill et al., 2017; Arslan, Bagchi and Ryu, 2015; Cagiltay, Cagiltay and Celik, 2020). Therefore, including similar data (e.g., future runs of the same course or, more broadly, further courses of the same discipline within the same platform), seems a wise strategy to expand the current certification predictive models. This is the next move towards improving some of the surveyed models; for instance, (Joksimović et al., 2016) found that more data, typically from the same subject domains as their current datasets, such as social science, should be analysed to account for diverse learning settings and more generalisable results. Their work is already diverse in terms of the dataset used, where the authors conducted an SNA based on learners’ social interaction in the discussion forum using two runs of the same course (one in English and one in Spanish). However, other replications of (Joksimović et al., 2016) intercultural study “but in different course disciplines” are still needed to reveal how learners’ interaction is different or similar based on their cultural backgrounds.

The above-suggested areas of future development correspond to the proposed future improvement mentioned in the studies surveyed. For instance, Tian et al. (2017) planned to collect more details about learner behaviours and\ examine the extent to which they may be more predictors of certification attainment. Also, Qiu et al. (2016) also found that more learner characteristics, or even newer courses, would be worth analysing using their proposed social network. In contrast, Arslan, Bagchi and Ryu (2015) plan to include other non-ICT-related features in the future, such as MOOC structure. These factors may influence certification prediction and help yield better model performance.

3.7.2. Sample extensive filtration

Preliminary filtration can be an essential step in data pre-processing for removing noisy unrepresentative data, such as removing logs of learners who never accessed the course and thus have zero logged activities. However, one noticeable drawback of some of the surveyed works is the extensive filtration of learners involved in the experiments, thus affecting the experiment generalisability discussed earlier and nominating a tiny sample of learners. Such misconduct is usually and intentionally exercised by including more data types (demographics, click streams, discussion forum activities, assignments, and quizzes).

Although MOOCs are developed and delivered in a similar approach, the comparison and generalizability of different models should at least be theoretically valid. Comparing the final outputs of the existing predictive certification models with the current research methods is not practical. Researchers tend to use highly-subsetted populations, which leads to the experiments being conducted under different conditions and the nature of the data; hence, the results would be risky to generalise. As discussed earlier, the studies reported different demographics and behaviours for being efficient in predicting certification, which is understandable due to the different data sources (platforms) and learner demographics. Therefore, population filtration, along with using many types of learner data (click streams, demographics, discussion forums, and surveys), negatively correlates with model generalisability. Also, there is massive diversity in the subpopulations analysed by the studies surveyed at the level of comparison. Even if two studies used the same type of data (e.g., question attempts or time spent), the other learner-related characteristics, such as demographics, are still unknown.

Although population filtration can help understand the relationship between a specific variable and the likelihood of certification, such as learners’ prior intention to obtain a certificate (Yeomans, Reich and Acm, 2017) or the uninterrupted time on task (Lee, 2018b), this deprives the experiment of broadly understanding the behaviours of far larger segments of the learners' population. Moreover, as shown in Table 3.2, the studies tended to lose many learners when more conditions (filtration) and data types were involved in the experiment.

We, therefore, believe that using one source of data (e.g., learner's demographics, forums discussions, question answering) might help build more generalisable models. This is especially the case with questionnaire-based experiments where the collected data are typically unique. For instance, surveying learners' intention of certification in Yeomans, Reich and Acm (2017) or their professional experience in years, daily work hours, English language skills, digital proficiency skills, previous experience in MOOCs, mother tongue, and average weekly available hours for study (Kostopoulos et al., 2021) are typically unique to these experiments or platforms.

3.7.3. Insufficient experimental elaboration

Some of the surveyed works, specifically ML-based experiments, have skipped reporting essential parts of their experiments, such as feature engineering and selection steps that are considered vital in building an efficient predictive model, but were nevertheless not commonly reported within the surveyed works. Parameter fine-tuning was less reported within the surveyed works than feature engineering and selection. With the exception of Kostopoulos et al. (2021), which was the only work that described in detail the parameters tuned for each employed predictive algorithm, the surveyed works seem to use their algorithm with the default values, depriving of many tuning benefits such as improving model performance, reducing training time, and countering over-fitting, especially on small datasets.

Another example of insufficient experimental elaboration is the random, unjustified approach followed. For instance, none of the surveyed works justifies adopting the employed predictive algorithms [i.e., whether (1) an initial experiment has been conducted to prove their performance on other algorithms or (2) they have been proved to outperform other algorithms according to the literature review]. Another example of an unjustified approach is the random training/test split. While 70:30 splitting is commonly followed during predictive modelling, some studies [e.g., (Ruipérez-Valiente et al., 2017) and (Tian et al., 2017)], which used 75% and 90% of the data, respectively, for training, did not justify why this specific splitting was decided and whether it improved model performance.

3.7.4. Non-realistic Modelling

One of the methodology-related concerns we noticed within the surveyed works is that some models were not realistically actionable in a real-life scenario. For example, while some works aimed at building timely intervention and early detection of learners' behaviours, others, marked as Early Model = N in Table 3.2, used up the whole course data for training the model, hence not applicable in an actual active MOOC, considering that learners decide on obtaining a certificate immediately after completing the course; hence, intervention at this late stage would be less practical. Another concern, even with the early predictive models, is training the model on early data [i.e., using the course's first week(s) and testing on the same course].

While this might be the only option for predictive models built only on one offering of the analysed course, multiple offerings-based models should temporarily build the experiment by training the model on all the early offerings of the course and testing on the last offering only. The latter concept was used to model other MOOC tasks, such as identifying at-risk learners (He et al., 2015) but was not followed by any of the certification prediction studies surveyed. However, it mimics real-life courses and would help build a more actionable, realistic predictive certification model. We understand that not all certification prediction works are aimed at real-time intervention; some studies have focused on explanatory and inferential analysis (i.e., statistical models). Thus, the issue discussed above is less relevant. However, for studies that promote real-time intervention and early prediction, the above concerns are essential to consider.

Having shed light on some limitations of current predictive models, it is essential to ensure that the limitations discussed above are simply out of the researcher's control. With most of these experiments being conducted on one course only, we understood that obtaining more data, either from subsequent runs of the same course or other courses, is challenging and restrains researchers from validating their models on larger datasets. We also considered researchers' challenges while attempting to access clean, complete, and modellable learners' datasets. However, these limitations are still valid for future model improvement.

The present thesis addresses several limitations identified within the surveyed literature review. Regarding generalisability, while the dataset we used was obtained from one platform only, it spans various runs (23) of 5 different MOOCs covering 4 distinct disciplines (literature, psychology, computer science, and business). This allowed us to longitudinally observe the changes in learners' activities from a paid certification perspective as further discussed in Section 5.3.1 and predict paid certification based on a rich source of data. As shown in Table 3.2, most of the surveyed works were based either on one course or run/iteration. Another novelty of the present thesis is predicting paid certification in MOOCs at an early stage (starting with using data from the first week of the course only). Our analysis showed that there are only 10 studies that adopted early prediction, out of which only one study explicitly predicts paid certification at an early point of the course using data from one run of an edX course. However, while edX has been extensively analysed (17 out of 25 studies), FutureLearn has never been explored from a paid prediction perspective, although it is considered the largest MOOC platform launched outside the United States.

The present thesis additionally presents a novel ensemble model for predicting paid certification based on data from discussion forums. This model synchronously processes textual (comments and replies) and numerical (number of likes posted and received, sentiments, POS tags) data from the forums and uses various DL algorithms for prediction, which was never done in the previous literature to the best of our knowledge.

3.8. Implications on Educational Practices

Keeping the recent proliferation of MOOCs in mind, the critical issue comes to light that the paid certification rate of a given course has also been declining over various runs/iterations; for instance, in some cases, the number of certificate purchases dropped by as much as 50% in the latest course run compared to the first run (Alshehri, Alamri and Cristea, 2021). This challenging and constantly low certification rate has prompted substantial research (Gitinabard et al., 2018) and pushed several providers to explore more sustainable business models (Dellarocas and Van Alstyne, 2013). In line with the efforts exerted to address this challenge, the surveyed studies on the prediction and decline of MOOCs paid certification introduced several implications for educational practices.

One of these implications is that the number of registered learners and course populations is declining, due to the transition of these platforms from semi-free to paywalled courses (Chuang and Ho, 2016). Despite the unparalleled success of MOOCs, especially in terms of the burgeoning learner enrolment, one of the more disturbing aspects to date is the staggeringly decreasing certification rates (Reich and Ruipérez-Valiente, 2019), a funnel with learners “leaking out” at various points along the learning pathway (Clow, 2013; Breslow et al., 2013). The issue is the balancing act between MOOC providers, trying to find ways to finance their offers, and MOOC consumers, the learners, who would prefer to learn for free (or very cheaply). From an educational perspective, *free* or *very cheap*, *high-quality*, *ubiqutous education* is a perfect way to reach the *largest coverage of the learners around the* *world*. Learners can learn via the *'any-time' and 'any-place' paradigm*, in some cases, in their native language, and still be receiving a world-class education.

However, the dropping numbers are concerning, both from an educational perspective – as learners are either not able or not willing to complete their courses; as well as from a financial perspective - as providers need to be able to explore course monetisation and platform sustainability via revenue generation. As has been explained in Chapter 2, there are many models for revenue generation. However, in this thesis, the focus is on the model that allows for cheap MOOCs, which support the ubiquitous, inclusive form of learning initially envisioned by the first MOOC proposers.

Another implication on educational practices is that, whilst from a business perspective, charging for certifications can provide a sustainable revenue stream for MOOC providers, it can at the same time provide funds that can be reinvested in *course quality, development, and platform maintenance*. Thus, paid certification can help platforms become more stable via revenue generation, but it also obligates providers to maintain *high-quality MOOCs* to attract paying learners. From an educational approach, this extends to having assessments being processed by humans and course materials regularly reviewed by experts, privileges some platforms currently provide exclusively for paying learners.

The surveyed studies in this chapter have also emphasised the importance of early intervention to address the lack of certification rate in MOOCs. However, although this thesis is focusing only on the certificate prediction, this does not need to be applied in practice by itself. Indeed, this can be combined with other approaches, including *personalising learners' experience*, as proposed by other research(Rohloff, Sauer and Meinel, 2020). Predictive analytics could help identify *individual learning styles*, *preferences*, and *needs* (Jena, 2018). Consequently, by understanding how students typically behave, platforms could tailor the content and learning experiences to match each student's learning style, which may, in combination, lead to a further improved certification rate (Gitinabard *et al.*, 2018). MOOC platform providers can consider implementing such adaptation methods in tandem, and educators may wish to recommend MOOCs based on the support they are providing to their learners. Educators may also wish to accept certain certificates provided by MOOCs they trust.

3.9. Epilogue

This SLR contributes to presenting a synthesis of the state-of-the-art studies on MOOC certification prediction, considering the struggle of MOOC platforms to build their own business models along with the recent transition, since 2017, towards paywalled content like micro-credentials, corporate training, and online degrees with affiliate university partners. It also serves as a roadmap for the multidisciplinary community of researchers in the educational domain (e.g., data scientists, statisticians, and educators) to explore the prediction of certification in MOOCs from a wider angle. We followed the 27-item preferred reporting items for systematic reviews (PRISMA) protocol for methodological rigour while conducting this SLR to increase the transparency and quality of the systematic review reported. The next chapter discusses the methodology followed for addressing the research questions stated earlier in Section 1.3.

Chapter 4 : Methodology

4.1. Prologue

Having highlighted the aims, objectives, and questions of the present research in Chapter 1; provided an overview of MOOCs in general and specifically certification in Chapter 2; and systematically reviewed the literature to identify and discuss the related works in Chapter 3; this chapter explains the methodology followed for answering our research questions, including the data collection and preprocessing, feature engineering and selection, statistical tests and classification approaches, and the performance metrics adopted for evaluating the developed models in Chapter 5, 6 and 7. Additionally, the ethical standards which were considered during writing this thesis were discussed in Section 4.6.

4.2. Data Collection

A dataset of almost 250,000 learners from 5 FutureLearn MOOCs, obtained by ourselves from Warwick University was used in Chapter 5. The dataset of Warwick University includes 23 runs spread over 5 MOOC courses on 4 distinct subjects. These topic areas are literature (course: Shakespeare and his World [SP], duration: 10 weeks); psychology (courses: The Mind is Flat [TMF], duration: 6 weeks; and Babies in Mind [BIM], duration 4 weeks); computer science (course: Big Data [BD], duration 9 weeks), and business (course: Supply Chains [SC], duration 6 weeks). These courses were delivered repeatedly in consecutive years (2013–2017). We used clickstreams in this experiment, whereas data extracted from learners' discussion forums, along with text sentiment annotations by MOOCSent, were used for conducting the third experiment, in Chapter 7.

One of the challenges of obtaining a MOOC dataset is that it is typically not owned by the platform. Course providers have the ultimate right to share learners' data with any third party, that is, the platform operates as an electronic medium to deliver the course. This fact makes it challenging for researchers because platforms such as FutureLearn have more than 250 partners[47].

In Chapter 6, we used a broader dataset, from several platforms, Coursera, Udemy, FutureLearn, Stanford, representing just over 1.2 million manually-annotated reviews and comments, to train and test MOOCSent. Learner reviews were used to train the model, whereas comments were used to test the models as the ultimate purpose of this experiment is to label the learner text inputs as accurately as possible with the author's sentiment. Table 4.1 presents the overall statistics of the datasets used in the three experiments conducted for the present thesis. Further explanations of the data collected and utilised in the three experiments are presented in Section 5.3.1, 6.4.1, and 7.3.1, respectively.

Table 4.1. Overall statistics of the datasets

Chapter	#Courses	Datatype	Sample Size
Chapter 5	5	Numerical	249,161
Chapter 6	633	Textual only	1,280,427
Chapter 7	5	Numerical & textual	28,638

4.3. Statistical Test

Our first step of exploring our dataset was examining whether it comes from a specific distribution. Kolmogorov–Smirnov test was conducted to ascertain this. As our data come from non-Gaussian (normal) distribution and the variables we are analysing are independent, we used the Mann-Whitney U test (also called Mann–Whitney–Wilcoxon (MWW) (McKnight and Najab, 2010)), a nonparametric test for testing the statistical significance of the difference of distributions. We use it

[47] https://www.futurelearn.com/partners

here to compare the activities of non-paying learners with certificate purchasers, as further explained in section 5.4.1.

4.4. Predictive Machine Learning Approaches

Since all the data utilised for training the baseline predictive/classifying models in the present thesis are appropriately labelled, supervised ML, a sub-field of artificial intelligence (AI), was adopted for predicting certification in MOOCs by employing algorithms to learn from samples (learner data), initially without applying certain programmed instructions (Goodfellow, Bengio and Courville, 2016). Various conventional ML, deep learning (DL), and NLP algorithms have been used through this study as detailed through the remaining subsections of this chapter. This section aims at presenting a theoretical background of the approaches followed to answer the research questions stated earlier. It also elucidates the technical concepts and the context that is essential to achieve the objectives of this thesis. Before explaining the algorithm used to conduct this research, we provide below a brief description of the various types of learning.

There are three different types of learning approaches in ML algorithms (supervised, semi-supervised and unsupervised), which can be adopted based on the nature of the problem and the available data for analysis. The supervision correlates with the data level of labelling, that is, unlabelled-data tasks tend to adopt unsupervised learning (such as clustering), semi-supervised learning is suitable for data with more unlabelled observations, whereas the fully (or the majority of observations) labelled-data tasks typically adopt supervised learning (such as prediction). The latter is known for its performance reliability (Alpaydin, 2020) and is therefore more commonly used compared to the other two learning approaches (Goodfellow, Bengio and Courville, 2016).

Within supervised learning, there are two main concepts of learning, shallow (or conventional), which requires human intervention for feature extraction and is typically done manually during feature engineering. One sub-concept of conventional ML is DL, wherein the algorithm automatically extracts the features. Figure 4.1 illustrates these two concepts of learning.

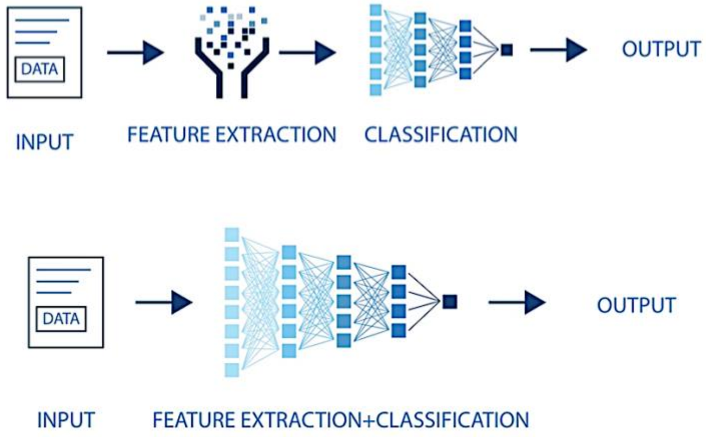


Figure 4.1. Comparison of learning concepts in ML (upper) versus DL (lower), cited from Khan *et al.* (2021b).

4.4.1. Conventional Models

Given below is a list of conventional models that are structurally simple and yet efficient in performance (Rosasco, 2016).

4.4.1.1. Logistic Regression (LR)

Logistic regression (also known as logit regression) is estimating the parameters of a logit model, which models the probability of occurrence of an event through its log-odds – a linear combination of the independent variables(s). LR is typically used for classification tasks but more commonly used for binary classification tasks (Cramer, 2002). It estimates the parameters of the coefficient in a linear combination in binary output classification as shown in Figure 4.2. The probability of the output (either certainly 0 or certainly 1) is estimated via logits (scales of log-odds as shown in Figure 4.2) which explain the name of the classifier (Shah et al., 2020; Rawlings, Pantula and Dickey, 1998).

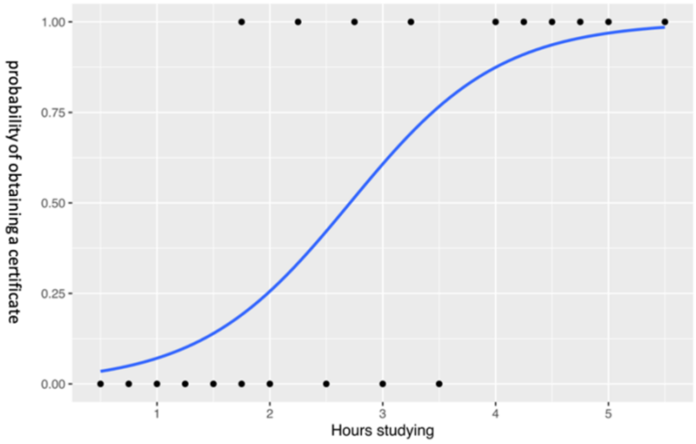


Figure 4.2. LR curve showing the (binary dependent variable) versus time spent on study in hours (continuous independent variable).

4.4.1.2. Decision Tree (DT)

Tree-based models are a set of classification/regression models where DT is the standard and the inspiration for the extended versions of the model, such as random forest and extremely randomised trees. Tree-based models repeatedly segment the input variables to build the DT using the most representative variables (features) of the dataset fed (Ali et al., 2012). The standard architecture of a DT consists of nodes and branches along with three building steps (splitting, stopping, and pruning) as shown in Figure 4.3, which illustrates an example of a single binary target (*y*) and two continuous variables (x^1 and x^2). Each node represents a choice which results in subdividing the records into subsets whereas the branches represent possible outcomes (occurrences) derived from a node, resulting in a hierarchy shaped DT. Hence, each node from the highest level of the tree (root nodes) to the lowest (leaf nodes) represents a rule of classification decision. Splitting controls which input variables are related to the target and is hence used for spitting a parent node into purer child nodes. The splitting procedure continues until the predetermined stopping criteria are met. The last component of a DT approach is pruning which aims first to grow a larger tree before pruning it (removing the less informative parts of the tree) until reaching an optimal tree size (Song and Ying, 2015).

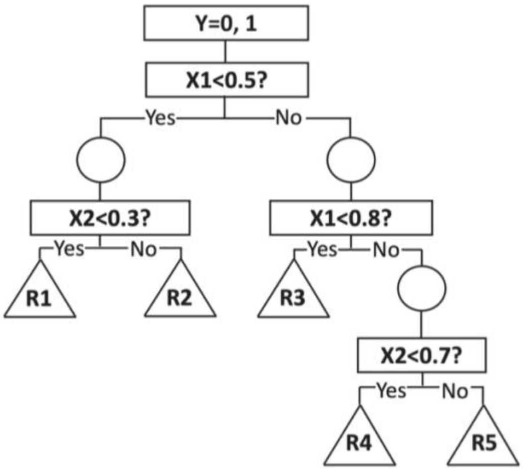


Figure 4.3. Example of DT based on a binary target variable, cited from Song and Ying (2015).

4.4.1.3. Support Vector Machine (SVM)

Support vector machine (also called support vector network) is widely used for classification, regression, and detection of outliers. SVM is primarily a non-probabilistic binary linear classifier (although some extended methods such as Platt Scaling use it as a probabilistic classifier, and it can be used as a non-linear classifier using Kernel Trick). SVM has been known for its competitive performance in high-dimensional space tasks. The algorithm maps training data as points in an n-dimensional space and then defines a hyperplane (which denotes the distance between the two groups using the mapped points (support vectors) as illustrated in Figure 4.4. The hyperplane is fitted where the maximum width of the gap between the two groups is. Finally, new examples are mapped into the margin and classified based on the side of the gap that is mapped. There are several estimators developed based on SVM including C-support vector classification (SVC), which was used in our experiments.

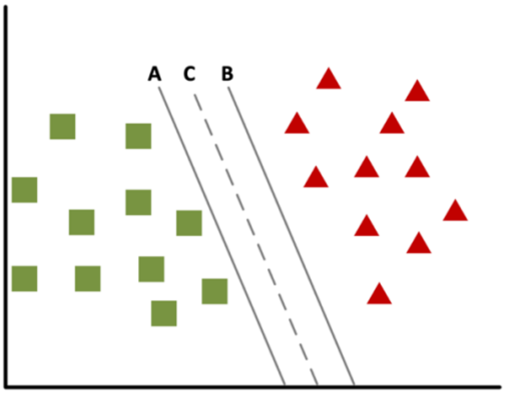


Figure 4.4. Hyperplane splitting and classification mechanism in SVM (C= optimal hyperplane, A-B= optimal margin), cited from Asharf *et al.* (2020).

4.4.1.4. Naïve Bayes (NB)

Naïve Bayes (or simple independent Bayes) is a Bayes' theorem-based probabilistic Bayesian classification technique. NB classifier assumes strong (naïve) independence between the features (i.e. the existence of one feature in a category has no relevance to the existence of other features). Regardless of its simplicity among other Bayesian network models, NB shows good performance in some sophisticated tasks such as multi-class and text classification tasks (Rish, 2001).

4.4.2. Ensemble Models

Ensemble modelling refers to using multiple diverse algorithms to predict an outcome. Therefore, it is based on the results of various classifiers that are combined and used for generating a majority vote out. While each individual algorithm has a different method of learning and depends on the application and the associated data for its performance accuracy, ensemble models can overcome this problem by achieving a generalisable result (being based on several voters (algorithms)) and at the same time addressing building a reliable model that can reduce variance and avoid overfitting (Asharf et al., 2020).

Ensemble learning can be based on bagging or boosting. Bagging trains n base learners based on random samples with replacement, allowing the model to proceed in parallel. In contrast, each model in boosting focuses sequentially on the misclassification of the previous model as illustrated in Figure 4.5.

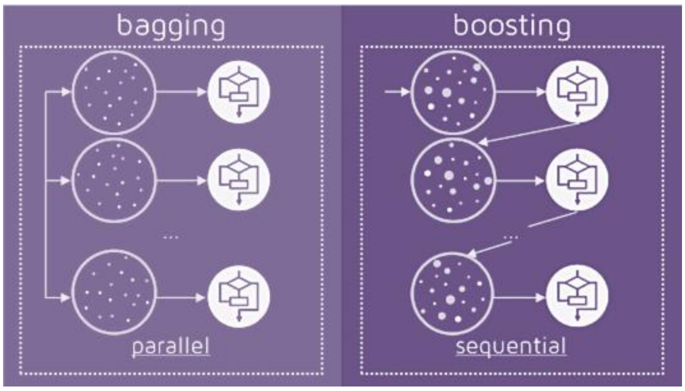


Figure 4.5. Bagging vs boosting ensemble learning methods, cited from Teja (2019).

In bagging, various DT are involved on different samples of the training data and then the prediction is averaged based on the results obtained from all the DTs involved. The best examples of this technique are random forest and extremely randomised trees. The second ensemble learning principle is boosting, which adds ensemble members sequentially so as to correct the predictions made by previous models (i.e. iteratively changes the input data to focus on the misclassified instances by the previously fitted models). Next, a weighted average of the predictions is calculated and provided as the output (Brownlee, 2021). Boosting-based ensemble models include adaptive boosting, gradient boosting machines, and stochastic gradient boosting.

4.4.2.1. Bagging-based Ensemble Models

4.4.2.1.1. Random Forest (RF)

RF is one of the tree-based (i.e. a forest of randomly created trees) algorithms that leverages the power of multiple DTs for classification. Thus, each node in the DT processes a random subset of the input variables to calculate the output. Finally, RF combines the outputs of each individual DT to make the final classification decision. Therefore, RF is an ensemble learning method that uses a combination of tree predictors where each tree depends on the value of a random vector. This algorithm samples these vectors independently and with the same distribution for all DTs in the forest (Breiman, 2001).

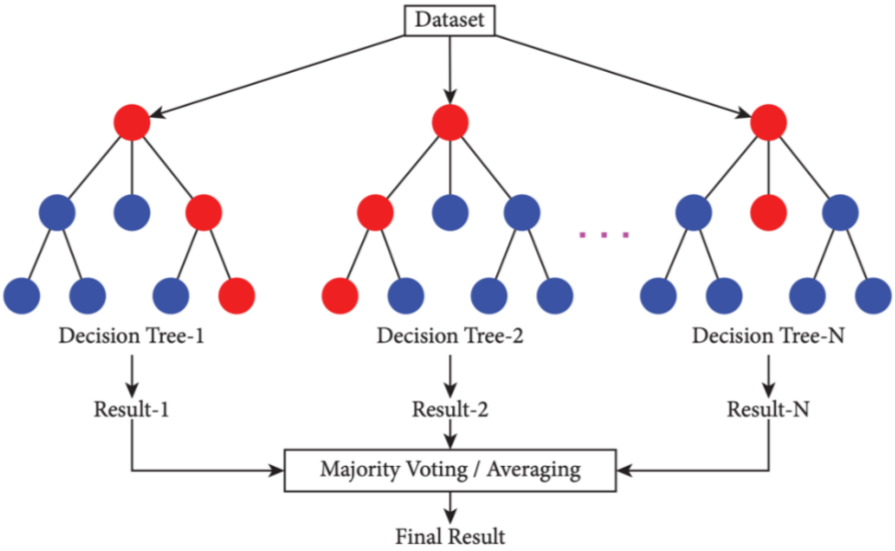


Figure 4.6. Illustration of RF classification technique, cited from (Khan *et al.*, 2021a)).

4.4.2.1.2. Extremely Randomised Trees (ET)

It is also known as Extra Trees (ET) and has a structure similar to RF, where both choose a random collection of characteristics for each node splitting but with two different mechanisms: sampling and node splitting. Compared to RF, ET samples from the entire dataset during the construction of the tree, whereas RF subsamples the inputs with replacement (bootstrap replicas). This allows ET to reduce the chance of bias as different subsets of data are more likely to introduce different biases. The other advantage ET has on RF is the randomised splitting of nodes within the DT, which leads to less influence by specific input variables and hence reduced variance. The improved structure of ET led to better performance on different benchmark datasets compared to RF (Geurts, Ernst and Wehenkel, 2006).

4.4.2.2. Boosting-based Ensemble Algorithms

4.4.2.2.1. Adaptive Boosting (AdaBoost)

AdaBoost is one of the early boosting-based ensemble algorithms which was introduced by Freund and Schapire to solve many complex problems, such as gambling, multiple-outcome prediction, and repeated games. It was based on the principle of combining the output of other learning algorithms (weak learners) to build a strong classifier (Freund and Schapire, 1997). As the name suggest, AdaBoost is adaptive to the misclassified predictions of the previous models, as it leverages and generates new weighted data point with less weighting assigned to the misclassified instances. Thus, the newly added model is more adaptive in addressing the previous model’s misclassifications (Schapire, 2013).

4.4.2.2.2. Gradient Boosting Machines (GBM)

Unlike AdaBoost which penalises wrong classification, GBM adopts the loss function (which can be the log loss in classification tasks or the mean average error in regression tasks) for improving the classification performance. Additionally, GBM continually minimises the loss, via utilising a gradient descent method, until it reaches the optimal point (Friedman, 2001).

4.4.2.2.3. Stochastic Gradient Boosting (XGBoost)

Stochastic gradient boosting is an optimised distributed GBM-based model with greater efficiency, flexibility, and portability. As a scalable end-to-end tree boosting system, XGBoost achieves better results with sparse data and weighted quantile sketch for approximate tree learning. XGBoost is a more regularised and extended version of GBM and hence has improved model generalisation capabilities (Chen and Guestrin, 2016).

4.4.3. Sentiment Classification Methods

4.4.3.1. TextBlob

TextBlob is an open-source text-processing Python library that allows one to perform several tasks, including noun phrase extraction, translation, part-of-speech tagging, sentiment analysis, tokenisation, and spelling correction. TextBlob is part of the well-known natural language toolkit (NLTK) and helps in reducing the computational cost of analysis. The tool generates a float value of a confidence level (between -1 and 1) for each text inserted and later annotates it as positive if > 0, negative if < 0, or neutral if = 0. These default thresholds however can be manually adjusted.

TextBlob assesses sentiment via returning a tuple of form (polarity, subjectivity, and assessments), where polarity and subjectivity float within a range of -1 and 1, with 0 being very objective and 1 being very subjective; assessments is a list of polarity and subjectivity scores for the assessed tokens.

4.4.3.2. VADER

Valence Aware Dictionary and Sentiment Reasoner (VADER) is a social media-based tool for general sentiment analysis. This open-source lexicon and rule-based tool uses a mix of qualitative and quantitative methods (a gold-standard list of lexical features along with their associated sentiment intensity measures), which are specifically attuned to sentiment in microblog-like contexts. Afterwards, the lexical features are combined, with consideration of five general rules, which embody grammatical and syntactical conventions, to express and emphasise sentiment intensity (Hutto and Gilbert, 2014). Similar to TextBlob, VADER generates a sentiment confidence level for each analysed text and allows resetting the thresholds of < 0, = 0, and > 0.

4.4.3.3. Stanza

Stanza is also an open-source Python natural language processing toolkit which can be used for lemmatisation, tokenisation, part-of-speech, multi-word token expansion, morphological feature tagging, sentiment tagging, dependency parsing, and named-entity recognition. This toolkit uses CNN for its architecture and massively supports more than 60 human languages. It was trained on 112 datasets, including the Universal Dependencies treebanks and other multilingual corpora. In comparison with the lexicon and rule-based tools, Stanza features a language-agnostic fully neural pipeline for text analysis, including a native Python interface to the widely used Java Stanford CoreNLP software. This makes it capable of more functionality and more advanced tasks, like relation extraction and co-reference resolution (Qi et al., 2020).

4.4.4. Deep Models

Deep learning is inspired by human brain neurons and can be defined as “a class of machine learning algorithms that uses multiple layers of nonlinear processing units for feature extraction and transformation” (Deng and Yu, 2014). Deep learning models are structurally based on artificial neural networks (ANN) that simulate human nervous system, which are connected to each other via electrochemical connections (synapses). Therefore, ANN follows the same structural design with a massive number of fully connected artificial neurons (AN) within adjacent layers. A weight parameter is assigned to each connection mimicking the neuron-to-neuron link in the biological synapse (Goodfellow, Bengio and Courville, 2016). The training step in a single neuron can be mathematically expressed as follows:

$$s = \sum_{i=1}^{n} w_i x_i \tag{4.1}$$

Where the weighted sum, *s*, is a neuron’s input calculated via multiplying the output of the previous layer’s neurons $x_1,x_2,\ldots x_n$ with their weights $w_1,w_2,\ldots w_n$. An activation function (e.g. *sigmoid*, *tanh* and *relu*) is then used to obtain the final output as illustrated in Figure 4.7 below:

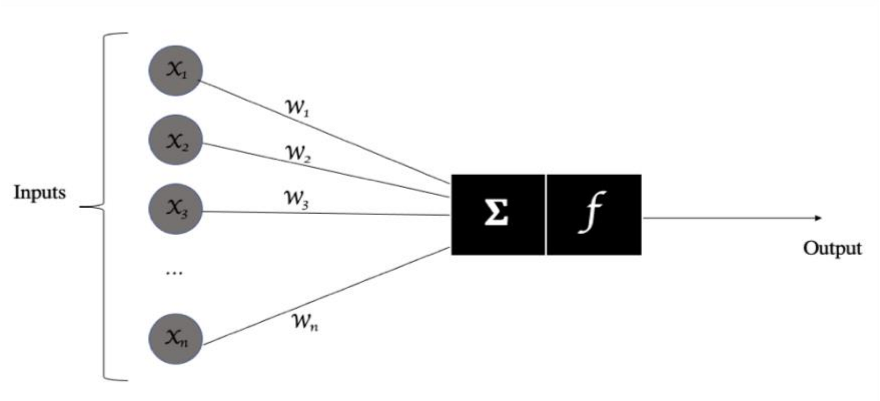


Figure 4.7. A single AN’s training process.

The final (output) layer, in classification tasks, uses an activation function (which adds a non-linear property to the output of a neuron, thus helping address complex problems) for generating a category probability distribution between the classes of the classification task. Next, a loss function is utilised to measure the gradient error values between the desired output (y) and the output ($\hat{y}$) generated by the ANN.

Training neural networks involves two main training steps. The first step is forward pass (layer-by-layer propagation) to generate output $\hat{y}$. The second step is backwoard pass (gradient values propagation) using the loss (cost) function backwards from the output layer. This function measures the deviation between the predicted output and the ground truth, based on which the NN weights are optimised. During the building of the neural networks, several parameters such as the learning rate, number of epochs, batch size, and the number of hidden layers can be chosen. These parameters significantly affect the network training process, where they can significantly improve the network performance if tuned properly (Bhardwaj, Di and Wei, 2018).

Overfitting is one of the common training issues an ANN may face, which occurs when the network becomes overly complex due to the high number of parameters (e.g. weights). There are various regularisation techniques such as dropout regularisation that can be used to control the risk of overfitting (Srivastava et al., 2014). Several DL techniques including CNN, RNN, LTSM, GRU, and BERT have been utilised in this study.

4.4.4.1. Convolutional Neural Network (CNN)

Akin to ANN, CNN (or ConvNet) consists of multiple weights and biases in layers but gives greater importance to the spatial structure of the input features. CNN has been commonly used for visual imagery, but it has also been effective in several NLP tasks such as search query (Chawla, 2021), sentiment analysis (Basiri et al., 2021), and text-based prediction (Kour and Gupta, 2022). CNN's standard architecture contains three layers: a convolutional layer (where the input data is converted into feature maps), pooling layer (where the feature maps are processed and parameter reduction takes place), and connected layer (where the output features are processed into a fully connected layer).

Unlike images, which are typically input as matrix of pixels in high-dimensional representation, textual data are input as a matrix of tokens of words or characters in a one-dimensional representation. There are some predetermined parameters (e.g. strides and filters) that are used to process such a 1D convolution via convolving the textual sequential input long vector (array) into a shorter vector.

In Figure 4.8, the input is learners' review which are tokenised into words, and each word is embedded into a word vector representation, thus an entire review is mapped into a matrix of size

$s \times d$, where s is the number of words in the review and d is the dimention of the embedding space (shown as 5 in Figure 4.8).

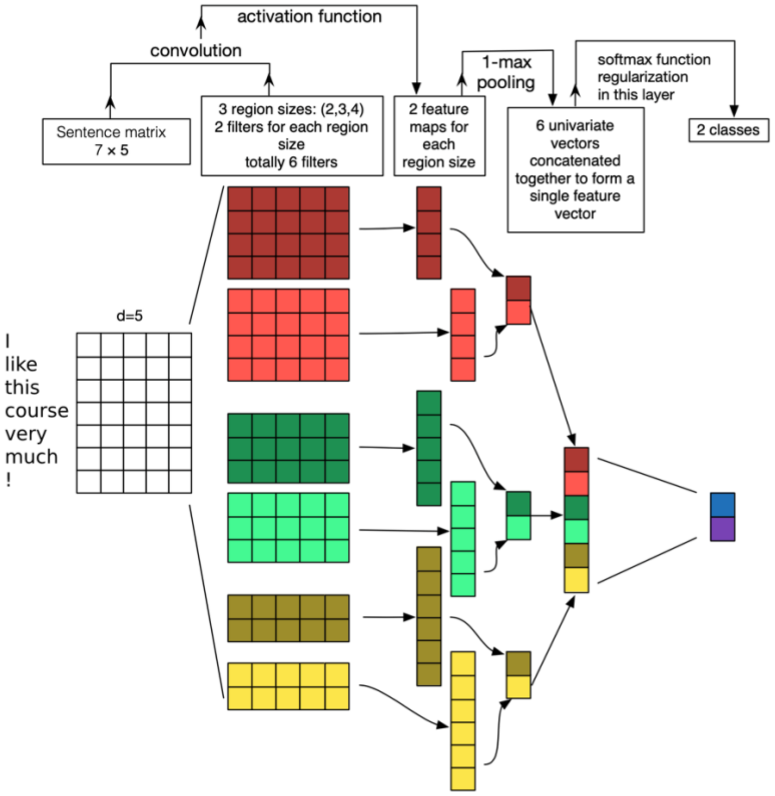

activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
course
very
much
!

Figure 4.8. CNN architecture for sentence classification, inspired by (Zhang and Wallace, 2015).

A convolution c_i is operated via applying the non-linear function f as follows:

$$c_i = f\left(\sum_{j,k} w_{j,k} \left(X_{[i:i+h-1]}\right)_{j,k} + b\right) \tag{4.2}$$

Where, c_i: the output (a concatenation of the convolution operator over all possible window of words in a learner review); $f(x)$: a non-linear function (e.g. relu); w: a weight; X: the current word embedding; i: the current input vector; h: the size of the convolution (#words spanned); j: a position in the convolution kernel/filter k; b: a bias term; and [i:i]: a submatrix of X.

Pooling layers are commonly used within CNN to reduce the size of the input while keeping the positional knowledge intact. *Max pooling* is one of the frequently applied pooling methods which passes the most important learnt features to subsequent layers while reducing the size (Sit et al., 2020). A max pooling is applied to each convolution, $c_{max} = max(c)$, to extract the most representative feature for each convolution, independently of where in the learners' review of this feature is located. Therefore, CNN is effective in identifying the most representative n-grams in the embedding space (Cliche, 2017). *Max pooling* is also used to generate a vector through combining all c_{max} of each filter, which first passes through a *fully connected hidden layer* and then through a *softmax layer* to compute the final classification probability. *Dropout layers* are typically added within the network for reducing the probability of model overfitting (Srivastava et al., 2014).

4.4.4.2. Recurrent Neural Network (RNN)

RNN is distinguished from feedforward neural networks (FFNN) in the way the inputs are handled, where the state of the RNN is not lost, as in ANN, after an input is processed. The output (values) of processing nodes in RNN is saved and fed back into the model recurrently; thus, each node acts as a memory cell and self-learns towards correct classification during backpropagation. All the states in RNN remain active throughout the sequence (acting as a memory), as the output of each current state is processed and concatenated with the input of the next step of the sequence. This style of learning is suitable for language modelling because texts are typically represented as a sequence of tokens (words or characters), which is the reason for adopting RNN and achieving outstanding performance in many NLP tasks (Yin et al., 2017).

RNN addresses the time twist, that standard FFNN has, through connections between passes and connections through time. A *hidden state* in RNN is calculated as follows:

$$h_t = f\ (h_{(t-1)}, x_t) \tag{4.3}$$

Where, x_t: the input vector x at time t and t-1: the *hidden state* of the previous input at time. Figure 4.9 shows the general architecture of RNN, where, x_t: input vector; h_t: hidden layer vector; o_t: output vector; and *tanh*: activation function.

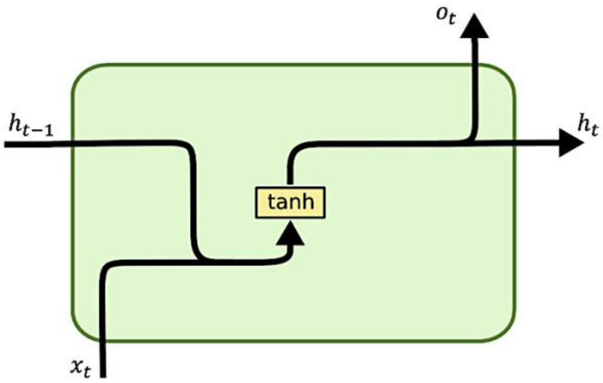


Figure 4.9. Architecture of RNN, cited from (lopez, 2019).

Backpropagation is another unique feature of RNN compared to ANN, which refers to the mechanism of minimising the loss function via adjusting the weights and biases of the RNN (updating weights and biases based on backward passes). The weight gradient is calculated based on multiple iteration (passes) and then summed. The new weight of a node is calculated as $w_{new} = w_{old} - \eta * \frac{\partial c}{\partial w}$, where η is the learnning rate and $\partial c/\partial w$ is the gradient of the loss function. Thus, the updated node weights depend on the gradients of the activation functions of each node in the RNN.

However, within large networks, backpropagation is associated with slowing the pace of RNN learning or, in some cases, not learning at all – a problem known as *vanishing gradient*, where the network weights remain unchanged. The gradient values are propagated to an earlier state and then either becomes smaller (progressively vanishes) or explodes, especially when modelling long dependencies such as lengthy sentences. This issue was tackled by developing LSTM, a more advanced architecture of RNN (Sherstinsky, 2020; LeCun, Bengio and Hinton, 2015).

4.4.4.3. Long Short-Term Memory (LSTM)

LSTM was introduced by Hochreiter and Schmidhuber (1997) as an extended version of RNN to control what information to send (or forget) to the next time step, which achieve unprecedented performance by being able to learn long-term dependency in data (LeCun, Bengio and Hinton, 2015). While each time step in RNN utilises one recurrent output only, LSTM produces a second recurrent output via a 4-gate cell as illustrated in Figure 4.10. The three gates in LSTM are as follows:

- Forget gate (for deciding which information needs attention and which to ignore): The current input x_t and hidden state h_{t-1} are passed via sigmoid function (σ) that generates a value of f_t between 0 and 1, measuring the importance of the previous output.
- Input gate (for updating the cell status): The current input x_t and previously hidden state h_{t-1} are passed through the second σ, which generate a value of i_t between 0 and 1, measuring the importance of the previous output. Next, the same x_t and h_{t-1} are passed through the activation function *tanh*, which generates a vector ($\tilde{c}_t$) with all possible values.
- Output gate (for determining the value of the next hidden state): The current input x_t and previously hidden state h_{t-1} are passed through the third σ. Next, the new cell state is passed through *tanh*. Thereafter, the final value is used by the network to decide on the information the hidden state should carry, which is later used for classification.

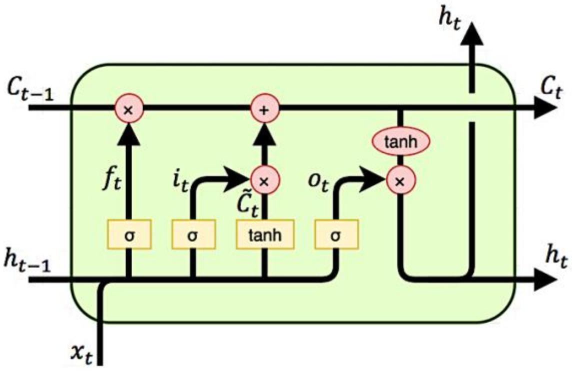


Figure 4.10. Architecture of an LSTM cell.

RNN gates can be mathematically expressed as follows:

$$f_t = \sigma\left(w_f \times \left[h_{(t-1)}, x_t\right] + b_f\right)$$
$$i_t = \sigma\left(w_i \times \left[h_{(t-1)}, x_t\right] + b_i\right)$$
$$\tilde{c}_t = tanh\left(w_c \times \left[h_{(t-1)}, x_t\right] + b_c\right) \tag{4.4}$$
$$c_t = f_t \times c_{(t-1)} + i_t \times \tilde{c}_t$$

$$o_t = \sigma\left(w_o \times \left[h_{(t-1)}, x_t\right] + b_o\right)$$

$$h_t = o_t \times tanh(c_t)$$

Where, *f* represents the forget gate of a cell, *i* the input gate, $\tilde{c}$ the new value for cell state, *c* the updated state value based on the sum of the products of old states $c_{t\text{-}1}$, *o* the output of σ which is utilised with the cell state c to compute the final decision.

4.4.4.4. Gated Recurrent Units (GRU)

GRU is similar to LSTM in terms of processing long-term dependencies. However, the distinguishing feature is that GRU merges both forget gate and input gate into one unit gate (update gate) as illustrated in Figure 4.11, thus previous memory is retained based on the size of the new dependencies (input). GRU consists of two gates (reset gate and update gate) and does not have protected hidden cell state, allowing full access to the allocated memory content (Chung et al., 2014).

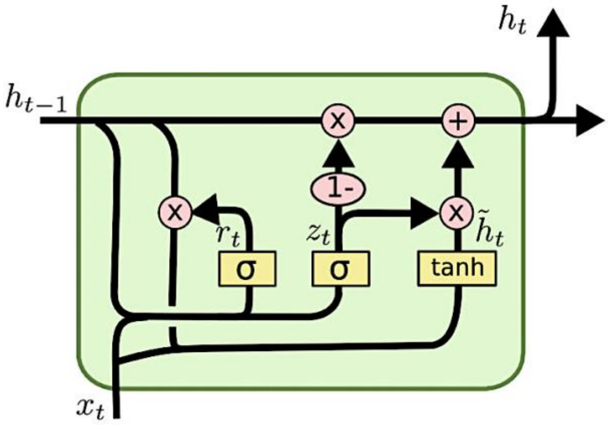


Figure 4.11. Architecture of GRU.

GRU can be mathematically expressed as:

$$z_t = \sigma\left(w_f \times \left[h_{(t-1)}, x_t\right]\right) \tag{4.5}$$

$$r_t = \sigma\big(w_r \times \big[h_{(t-1)}, x_t\big]\big)$$

$$\tilde{h}_t = tanh\big(w_c \times \big[r_t \times h_{(t-1)}, x_t\big]\big)$$

$$h_t = (1 - z_t) \times h_{(t-1)} + z_t \times \tilde{h}_t$$

Where z represents the update gate, r reset gate, which are calculated similarly to the input gate and forget gate in LSTM but without adding them to the formulas (first and second above). Additionally, LSTM changes the current hidden layer h, whereas in GRU the input x and the previous layer $h_{(t\text{-}1)}$ can modify the update gate and reset gate values. finally, the current layer is updated by z and r accordingly.

4.4.4.5. Transformers

Innovative DL models have been proliferating, and new models are regularly being introduced in recent years. One of the recently introduced neural network models is Transformer (Vaswani et al., 2017; Wolf et al., 2019), which has been trending specifically within NLP tasks. Unlike other NLP systems that are RNN-based in terms of structure, Transformers are based on attention mechanism and encoder-decoder structure (Cho et al., 2014; Vaswani et al., 2017). Similarly to RNN-based models, the input data are sequentially handled by transformers; nevertheless, the essential difference is that transformers do not necessarily handle the input data in order given that its mechanism recognises the context and meaning each word represents within a text. This makes transformers "the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution" (Vaswani et al., 2017)

4.4.4.5.1. BERT

BERT is one of the architectures that are backed by transformers and mainly aims for understanding the relationship between words. It is one of the most advanced language representation models for a broad range of NLP tasks, such as question answering, language inference, and sentiment analysis. BERT is developed via pre-training a deep bidirectional representation, by jointly conditioning a two-way context for all layers. Unlike directional models (left to right or right to left) such as RNN and LSTM, BERT and all other transformers process the whole sentence as one input rather than sequential ordering, resulting in contextual learning of each word with respect to all other words in the sentence (Jose, 2020).

BERT has two parameter-intensive settings: (1) BERTBASE: 12 layers, 768 hidden dimensions and 12 bidirectional self-attention heads with 110 million parameters and (2) BERTLARGE: 24 layers, 1,024 dimensions and 16 bidirectional self-attention heads (in transformer) with 350 million parameters. BERT is trained from unlabelled data obtained from Wikipedia (2,500M words) and BookCorpus (800M words) to be fine-tuned for any NLP task (Devlin et al., 2018).

The out-of-box version of BERT is excellent for general NLP tasks, whereas further training with masked language modelling (MLM), which is one of the training mechanisms adopted by BERT, may work better for domain-specific tasks (Briggs, 2021). With MLM training approach, 15% of the words in a sequence is masked with [MAS] tokens. For example, a learner comment that contains the text “this course is amazing” would be represented as “this course is [MASK]” with a vector representing each word. Next, the model will predict the masked part of the sentence with respect to all other words (Jose, 2020).

As illustrated in Figure 4.12, the encoder output of the transformer encoder is passed to a fully connected classification layer. Next, the output is multiplied by an embedding matrix to calculate the predicted probability of the words. The loss function only handles the predictions of the masked words. BERT has two procedures, namely pre-training and fine-tuning. The same architecture is followed in both, apart from the output layers as shown in Figure 4.12. In pre-training, the same model parameters are used for initialising the models for different tasks, whereas in fine-tuning all the model parameters are fine-tuned.

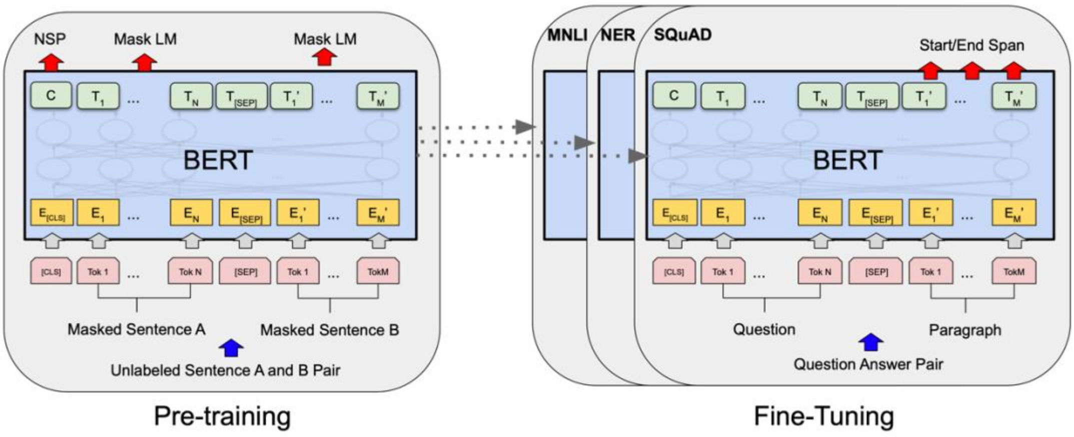


Figure 4.12. BERT procedure for pre-training versus fine-tuning. [CLS]: special symbol added in front of every input example; [SEP]: special separator token, cited from (Devlin *et al.*, 2018).

Another BERT training mechanism is next sentence prediction (NSP). During training, the sentences are received in pairs, with half of them having exactly the same consecutive sentence as the second term, and the rest having random sentences as the second sentence. BERT eventually predicts the actual second sentence (whether actually connected to the first sentence) after applying some embedding procedures before feeding the sentences to the models as shown in Figure 4.13. The input embeddings are typically the sum of the three embeddings of tokens, segments, and positions.

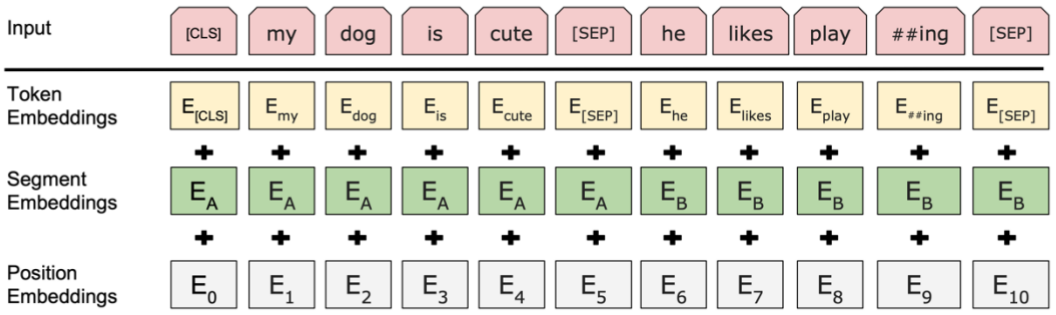


Figure 4.13. BERT input representation, cited from (Devlin *et al.*, 2018).

4.5. Performance Metrics

Several performance metrics have been used to measure the performance of the models developed within this thesis. This includes the standard metrics of recall and precision which can be computed from the confusion matrix illustrated in Figure 4.14.

		Predicted Classes	
		Positive	Negative
Actual Classes	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 4.14. Confusion Matrix of performance measurements.

While recall measures how well a model predicts the positive instances, precision calculates the ratio of positive predicted instances that are in fact true positive instances. The mathematical expressions of these two metrics are:

$$Rec = \frac{TP}{TP + FN} \tag{4.6}$$

$$Prec = \frac{TP}{TP + FP} \tag{4.7}$$

These metrics measure a model's performance at the level of class. Therefore, for the model's overall performance, we have further adopted other metrics such as F1, balanced accuracy (BA) and the total accuracy (Acc). F1 can be considered a trade-off between precision and recall. BA is defined as the average of recall obtained on each class which equals the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate). Acc. calculates the model's overall performance, giving the same weighting for all classes by dividing the total number of correctly predicted instances by the total number of instances. Given below are the mathematical expressions for these performance metrics.

$$F1 = 2 * \frac{Prec * Rec}{Prec + Rec} \tag{4.8}$$

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FN}\right) \tag{4.9}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.10}$$

4.6. Ethical Consideration

Predictive modelling of learners' behaviours in MOOCs has various positive and potentially beneficial applications to make learner's experience better. This research is typically conducted using the already collected and extracted learners' data. A consent from the learners to use their data for research purposes, typically during registration, is already obtained before this. However, this does not consequently grant a researcher the right to handle the data before some ethical standards are strictly considered. This includes researching towards the explicitly positive outcomes of the research, that is, not using learners' data, especially sensitive ones such as financial data used in our analysis, for purposes other than those stated within the research. Other research ethical standards, referred to by the university ethics committee, include keeping the data secure and anonymising learners' data – prior to analysis – by removing data that offer immediate identification of learners. All these standards were followed while undertaking the present research.

4.7. Epilogue

This chapter explains the methodology followed in answering the research questions. This includes description of the datasets collected from various MOOC platforms, the data preprocessing followed, features selection, statistical tests, predictive ML approaches and the metrics used to evaluate the models' performance. In the next chapter, we use learners' weekly activities to predict their purchasability of the course certificate.

Chapter 5 : Predicting Paid Certification in MOOCs Based on Learners' Weekly Activities

5.1. Prologue

This chapter focuses on using learners' weekly activities (accesses, attempts, and correct and wrong answers) and the time spent on each learning step to predict paid certification in MOOCs. The experiment is based on 23 runs of 5 MOOCs obtained from the relatively unexplored MOOC platform of FutureLearn to first statistically measure the difference in terms of activities between non-paying learners and course purchasers and then predict course purchasability using various ML classifiers.

5.2. Introduction

Massive open online courses (MOOCs) have been more population due to the free or low-cost offering of their content since their noticeable emergence in 2012. While this has helped these platforms to gather millions of learners in just a decade, the certification rate of both free and paid courses has been declining. Thus, this chapter uncovers the latent correlation between learner activities and their decision to purchase a course certificate via statistically comparing the activities of non-paying learners with course purchasers and predicting course certification using different classifiers, optimising for this inherently strongly imbalanced dataset. Our results show that learner activities are the best predictors of course purchasability using only the learner's number of step accesses, attempts, correct and wrong answers and time spent. achieving promising BAs. across the 5 courses.

Clickstreams have been adopted for prediction in MOOCs. Wintermute, Cisel and Lindner (2021) conducted a network-based exploration of learners' achievements by examining how a course-course interaction affects the likelihood of certification, and found that user engagement positively correlates with the certification rate in all 140 courses analysed. Wang, Hemberg and O'Reilly (2019) studied the impact of learner-obtained grades on their activities during the remaining content of the course and found that learners' behaviours did not change significantly after reaching the minimal grade for certification. Additionally, Coleman, Seaton and Chuang (2015) used LDA to explore behavioural patterns via learners' clickstreams to predict the likelihood of certification, whereas Ruipérez-Valiente et al. (2017) used learners' assignments, visits, and time spent on various activities for the same prediction target.

These statistical and ML-based models, and other models that are explained in Section 3.5, which were based on other types of MOOC data, are either were based on one course or run/iteration, did not consider weekly (early) prediction or use data of *paid* courses. Considering the recent transition of MOOCs towards paid macro-programmes and online degrees with affiliate university partners, this chapter presents a fine-grain exploration of learner behaviours from a different point of view – non-paying learners versus certificate purchasers. Specifically, this chapter attempts to answer the following research questions:

- *RQ1: Do non-paying MOOC learners behave differently from course purchasers as to their activities of access and answering questions (attempts, correct and wrong answers)?*
- *RQ2: Can MOOC learner's clickstream data (accesses, attempts, correct and wrong answers) and time spent on course steps predict paid certification for courses?*

It is worth mentioning that the first research question attempts to compare the activities of non-paying learners (NL) versus certificate purchasers (CP) using a systematic statistical methodology as shown in Section 3.5. Subsequently, the second research question examines whether learners' activities can be used to predict later certification behaviour.

5.3. Experimental Setting

5.3.1. Data Collection

When a learner joins FutureLearn for a given course, the system generates logs to correlate unique IDs and time stamps to learners, recording learner activities such as weekly-based steps visited, completed, comments added, or question attempted (Alshehri et al., 2018). The current study is analysing data extracted from a total of 23 runs spread over 5 MOOC courses, on 4 distinct topic areas, all delivered through FutureLearn, by the University of Warwick. These topic areas are literature (course: Shakespeare and his World [SP], duration: 10 weeks); psychology (courses: The Mind is Flat [TMF], duration: 6 weeks; and Babies in Mind [BIM], duration 4 weeks); computer science (course: Big Data [BD], duration 9 weeks), and business (course: Supply Chains [SC], duration 6 weeks).

These courses were delivered repeatedly in consecutive years (2013–2017); thus, we have data on several "runs" of each course. Table 5.1 shows the number of enrolled, non-paying learners (NL), as well as those having purchased a certificate (CP). Our data shows that learners accessed 3,007,789 materials in total and declared 2,794,578 steps completed. Regarding these massive numbers, Table 5.1 clearly illustrates the low certification rate (less than 1% of the enrolled learners).

Table 5.1. The number of non-paying learners and certificate purchasers on 5 FutureLearn courses.

Course	#Runs	#Weeks	#Steps	#Non-paying Learners	#Certificate Purchasers
BD	3	9	105	33,427	265
BIM	6	4	75	48,771	670
SC	2	6	118	5,808	69
SP	5	10	134	51,842	500
TMF	7	6	93	93,601	314
Total	23	35	525	233,449	1,818

While the very low certification rate has already been threatening MOOCs sustainability, subsequent runs yield worse certification rates. Our analysis shows that courses tend to lose enrollees and, consequently, purchasers, over the consecutive runs of a course, as shown in Table 5.2.

Table 5.2. The number of non-paying learners and certificate purchasers in each run.

Course/Run	**#Non-paying Learners**	**#Certificate Purchasers**	**%**
BD	33,427	265	0.79
BD1	16,385	118	0.72
BD2	11,281	75	0.66
BD3	5,761	72	1.25
BIM	48,771	670	1.37
BIM1	12,651	185	1.46
BIM2	9,740	168	1.72
BIM3	7,765	104	1.34
BIM4	6,225	78	1.25
BIM5	8,443	85	1.01
BIM6	3,947	50	1.27
SC	5,808	69	1.19
SC1	4,572	50	1.09
SC2	1,236	19	1.54
SP	51,842	500	0.96
SP2	15,914	227	1.43
SP3	12,692	111	0.87
SP4	15,881	102	0.64
SP5	7,355	60	0.82
TMF	93,601	314	0.34
TMF1	10,058	29	0.29
TMF2	22,929	48	0.21
TMF3	15,068	98	0.65
TMF4	10,314	43	0.42
TMF5	13,463	37	0.27
TMF6	14,254	37	0.26
TMF7	7,515	22	0.29
Grand Total	233,449	1,818	0.77

5.3.2. Data Preprocessing

The obtained dataset went through several processing steps so as to prepare them before feeding them into the learning model. Considering some learners were found to be enrolled on more than one run of the same course, the run number was attached to the learner's ID, to avoid any mismatch during joining learner activities over "several runs" with their current activities.

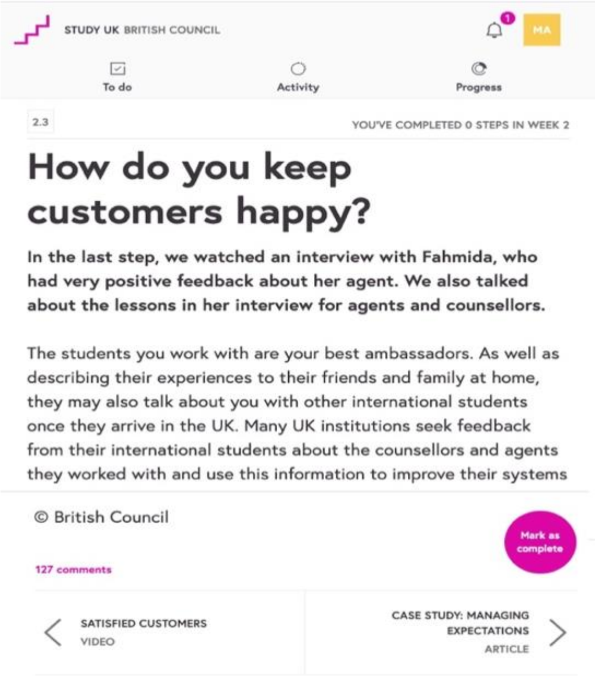
The preprocessing further contained eliminating irrelevant data generated by organisational administrators (455 admins across the 23 runs analysed). Table 5.3 shows the four main features analysed in this study.

Table 5.3. The features utilised for comparing learner activities and predicting course purchasability.

Activity Source	Activities (per week)
Step Access (a)	#Accessed steps
Attempts (t)	#Attempts
Correct Answers (r)	#Correct Answers
Wrong Answers (f)	#Wrong Answers

5.3.3. Time-spent Analysis

As the number of the step access was found to be a significant feature among our dataset, we further extended our analysis to examine the extent to which the time spent by learners on each step can predict paid certification. The step-based time spent feature, tspent –which we used here for prediction and which proved to be a highly representative factor helpful for learners' purchasing prediction – represents a computed value (rather than being provided as a log variable within the obtained dataset). This feature was defined as the difference between the first time a given learner accesses a step (first_visited time stamp) and the time when that step is fully completed (last_completed time stamp), as per learner's declaration (by clicking the 'Mark as Completed' button), as shown in Figure 5.1.



2.3

YOU'VE COMPLETED 0 STEPS IN WEEK 2

How do you keep customers happy?

In the last step, we watched an interview with Fahmida, who had very positive feedback about her agent. We also talked about the lessons in her interview for agents and counsellors.

The students you work with are your best ambassadors. As well as describing their experiences to their friends and family at home, they may also talk about you with other international students once they arrive in the UK. Many UK institutions seek feedback from their international students about the counsellors and agents they worked with and use this information to improve their systems

© British Council

127 comments

Figure 5.1. The interaction component on a Weekly *step* Page

The time spent can be mathematically expressed as:

$$ts_{(l,s)} = tc_{(l,s)} - tv_{(l,s)} \tag{5.1}$$

where *ts*=time spent, *tc*= last_completed time stamp, *tv*= first_visited time stamp, *l*=the current learner, *s*=the current step.

As our dataset has several logged dates and respective times for various activities of the learners in the system, the pd.to_datetime function was applied, to convert these variables into a set of strings (year, month, day, hour, minute) to enrich our input features and allow for an as high performance as possible with the few features available, as well as taking into account the aim to use features available early in each run to allow for early predictability. The latter aspect is critical as it means that course providers could use our prediction model to create early interventions and thus guide more of the learners towards paying behaviour, potentially increasing their revenue.

5.3.4. Feature Extraction

The preliminary data shape is a timestamp log spread on different data frames based on the data log source (access log, question answering log, and time-spent log). As MOOCs are usually delivered on a weekly basis, it was essential to compute the various weekly activities of each learner generating a temporal matrix of their weekly activities. The newly processed learner activities matrix of each course is as follows:

$$la = \begin{bmatrix} l_1 & a_{w(1-n)} & t_{w(1-n)} & r_{w(1-n)} & f_{w(1-n)} & ts_{w,s(1-n)} \\ l_2 & a_{w(1-n)} & t_{w(1-n)} & r_{w(1-n)} & f_{w(1-n)} & ts_{w,s(1-n)} \\ l_n & a_{w(1-n)} & t_{w(1-n)} & r_{w(1-n)} & f_{w(1-n)} & ts_{w,s(1-n)} \end{bmatrix}$$

Where, *l* = learner ID (excluded during building the model), *a* = access, *t* = attempts, *r* = correct answers, *f* = wrong answers, *w* = week, *n* = the number of the weeks in a given course, s = step(s) in week *w*.

5.3.5. Features Selection

Our preprocessed number of features as can be seen in the la matrix above is considerably high because the total number of the main extracted features (4) has to be multiplied by the total number of weeks w in a given course c, in addition to the computed weekly time-spent features. This resulted in a large array of features, especially for long courses such as SP, where the number of weeks was 10, generating 174 features (40 features of access, attempts, correct and wrong answers, and 134 times pent-based features). However, while this would allow (1) a temporal fine-grain analysis of the course content and (2) a timely and early prediction of learner's behaviours, in order to highlight the most representative features, feature selection techniques were applied.

One of the key steps of data preparation for conventional ML models is feature selection which nominates the most representative variables, eliminates any none-informative (irrelevant) variables, helps build a more interpretable and less costly model, and more importantly, improves the performance of the classifier (Kuhn and Johnson, 2013). Also, considering our data comprise numerical inputs (predictors) and categorical output, Kendall's Tau (Kendall rank correlation coefficient) was applied to measure the correlation between our predictors and target. Given below in (Figure 5.2-5.6) is a per-course illustration of the most important (correlated with the output) features truncated to 30 features due to page width limit. These are the most important features over the entire course; however, the features selected for early prediction (using only the first week data and the first half of the course) were selected accordingly.

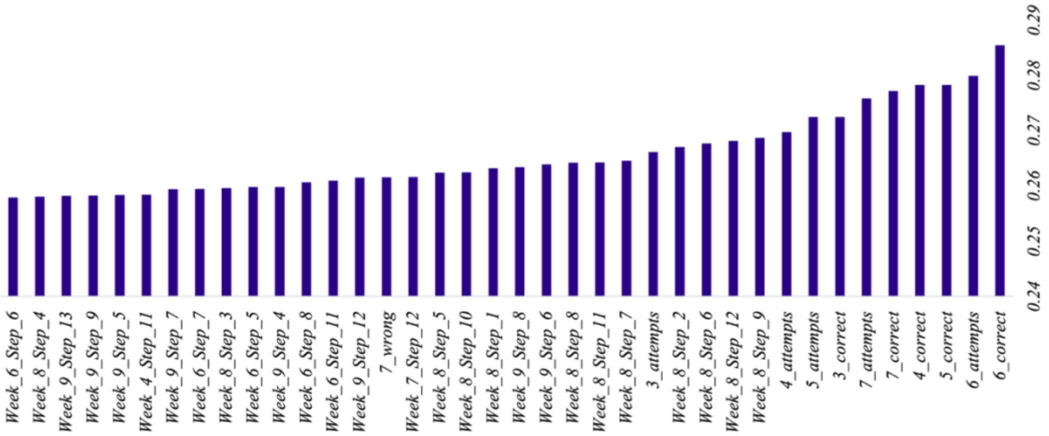


Figure 5.2. Feature selection in BIM course.

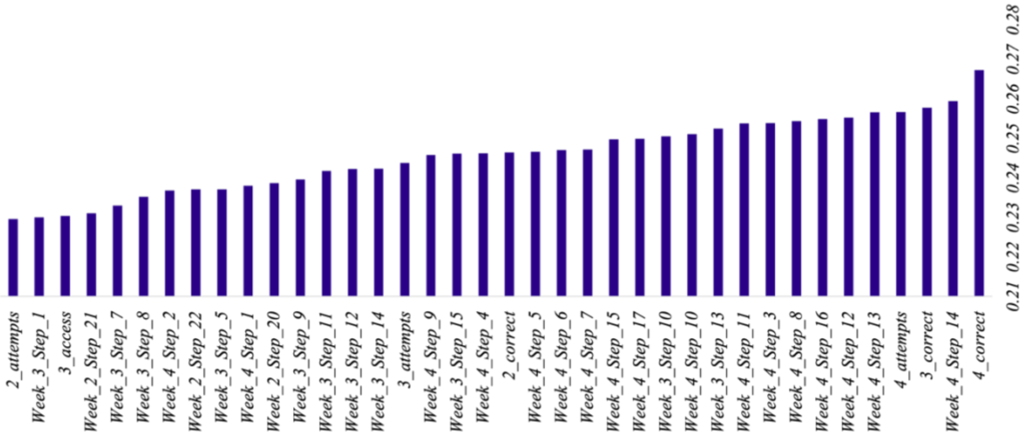


Figure 5.3. Feature selection in BD course.

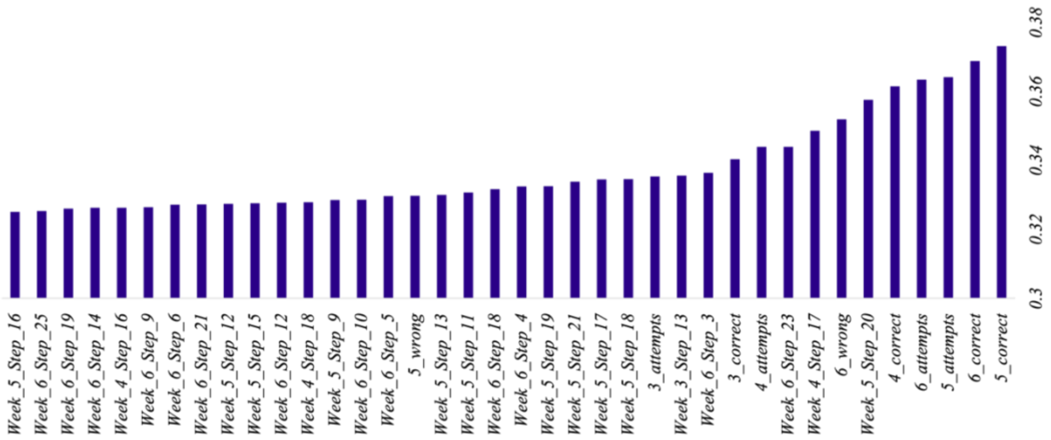


Figure 5.4. Feature selection in SC course.

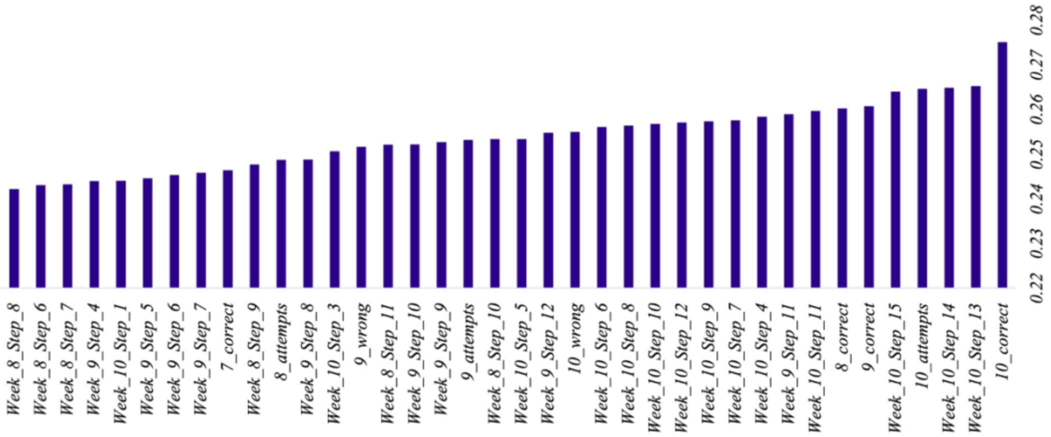


Figure 5.5. Feature selection in SP course.

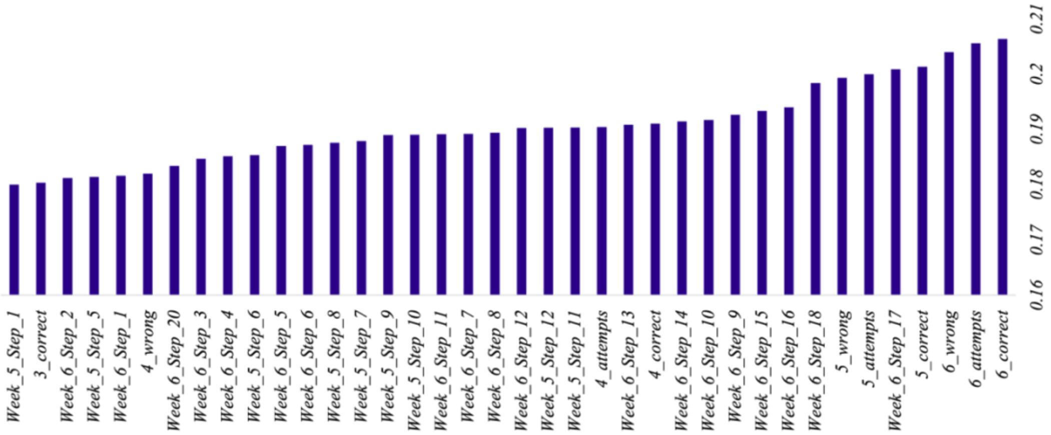


Figure 5.6. Feature selection in TMF course.

5.3.6. Statistical Analysis

Our first step in exploring our dataset was examining whether it comes from a specific distribution. To ascertain this, the Kolmogorov–Smirnov test was conducted. As our data come from non-Gaussian (normal) distribution and the variables we are analysing are independent, we used the Mann–Whitney U test (also called Mann–Whitney–Wilcoxon (MWW) (McKnight and Najab, 2010)), a nonparametric test for testing the statistical significance of the difference of distributions. We use it here to compare the activities of non-paying learners with certificate purchasers. The *U* value can be mathematically expressed as follows:

$$U = R - \frac{n(n+1)}{2} \tag{5.2}$$

Where, *n*: sample size and *R*: the sample's sum of the ranks.

5.3.7. Classification Algorithms

Further to the statistical inference, the current study applied four different classification and regression algorithms to predict MOOC learners' purchasing behaviour: RF, ET, LR, and SVC. These algorithms were chosen because they were able to predict course purchasability well by dealing with the massively imbalanced datasets and using at the same time only a very few features, as shown in Table

5.3. These input features exist in any standard MOOC system, which further promotes our model as generalisable. There are some further features that can be utilised for learner behaviour prediction, for example, demographics or leaving surveys; these features are either not generated by every MOOC platform or logged later after the end of the course, making the early prediction of purchasing behaviour challenging.

To simulate the real-world issue of the low certification rate in MOOCs, we fed the imbalanced data to the classification models without any modifications. We have initially used many other classification algorithms for this prediction tasks. However, the algorithms that do not deal well with imbalanced data, that is, those that have a parameter to define the class weight during learning, were excluded.

To deal with our imbalanced dataset, we used the BA metric, which is defined as the average of recall obtained on each class (developers, 2007-2020). BA equals the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate) as follows:

$$BA = \frac{1}{2}\left(\frac{tp}{tp + fn} + \frac{tn}{tn + fn}\right) \tag{5.3}$$

5.3.8. Dealing with Bias

Algorithms have been playing a significant role in shaping various aspects of e-learning systems, and consequently, many view their decisions and predictions as inherently objective and fair (Lee, 2018a) (Lee, 2018). However, a contrasting viewpoint has recently arisen, suggesting that algorithms frequently incorporate the prejudices of their developers or the broader society, leading to predictions or conclusions that exhibit evident bias against particular groups of learners (Baker and Hawn, 2021). Bias in education has been documented since the 1960s, with a recent high observation in predicting learners' success and failure in schools (Christie *et al.*, 2019) or, more importantly, within MOOC predictive models (Gardner, Brooks and Baker, 2019).

The literature shows that algorithmic bias in education is heavily associated with learners' demographics, namely gender identity, race, nationality, ethnicity, age, national origin and sexual orientation (Baker and Hawn, 2021). In our dataset used in the current experiment, the learners in each course have been included, regardless of their characteristics or demographical information. This is not only due to learners' demographics missing from our dataset, but also for developing generalisable models for predicting paid certification. This is in line with avoiding extensive sample filtration, which has been identified as one of the common limitations in the previous predictive model, as further explained in section 3.7.2.

Another common approach to mitigate potential bias is data preprocessing and algorithmic metrics (Baker and Hawn, 2021), which were adopted in this experiment, to enhance the reliability and fairness of the predictive models. This includes data shuffling before training, which ensures that the predictive model is exposed to all different types and patterns of learners' activities while learning. As a result, this strategy ensures the model remains general and does not overfit the training data through training on specific groups of learners. We also adopted the stratified cross-validation (CV) technique, which uses *k* folds (portions) of the data, preserving the same percentage of samples for each class in each fold to train and test the model on different iterations. CV architecture is commonly known for avoiding overfitting, by allowing the model to train on multiple train-test splits. Consequently, a better estimation of the model performance on unseen data is ensured. With *k*=10 and the statistics in Table 5.1 in mind, the training and test data size ratios are 30322:3370, 44496:4943, 5289:586, 47107:5235 and 84523:9390 for the courses BD, BIM, SC, SP and TMF, respectively.

5.4. Results

5.4.1. Statistical analysis

The results explore how our processed features can temporally identify course buyers based on their activity data. Our temporal analysis showed some statistical significance at various levels when comparing the behaviours of non-paying learners with those of the certificate purchasers across the 5 courses analysed. Tables 5.5-5.8 show the statistical analysis results where **bold** values mean the most significant value in each course. As the courses analysed spanned over different numbers of weeks, we have selected the first, middle, and last weeks to report the results. Below are the results for the four analysed activities (Access, Attempts, Correct Answers, and Wrong Answers).

For courses with an even number of weeks, we have selected the middle week closer to the start of the course for analysis. As shown in Table 5.5 to 5.8, our analysis indicated that paying learners were generally more engaged with the course content in accessing the content more frequently, attempting more questions, answering more questions correctly, and reattempting more questions answered incorrectly. Since platforms allow learners several attempts to answer a question correctly, the latter (number of wrong answers) indicates certificate purchasers' persistence to reach the minimum score required to be eligible for the course certificate. While all the statistical analysis results were very significant ($p < 0.001$, ranging from 4e-23 to 0), the test showed higher prediction power towards the end of the course. Course-wise, the difference in the activities of both groups of learners in SH was the most significant, whereas the significance of these four predictors based on the results of the last week can be placed in descending order: as attempts, correct answers, wrong answers, and access. Thus, non-paying course takers behave differently from course purchasers, in accessing content and answering questions (attempts, correct answers, and wrong answers).

5.4.1.1. Access

Table 5.4 below shows the statistical analysis results of comparing the number of accesses for non-paying learners and certificate purchasers at three different time points of the course. **Bold** values refer to the most significant value in each scenario (week) in each course, whereas *italic* values refer to the *p-value*. Since the courses analysed spanned over different weeks, we have selected the first, middle, and last weeks to report the results. As seen in Table 5.5, the difference in the number of accesses becomes larger when learners reach the end of the course, with the last week generally having the most significant results. The only exception is the BD course, where the mid-week activity difference was more significant compared to the first and last weeks. All the statistical analysis results were significant, with *p* values ranging between 4e-23 and 6e-264 for SC 1st week and TMF last week, respectively.

Table 5.4. Comparison of the number of accesses for non-paying learners and certificate purchasers at three different time points of the course.

Course	Measure	1st Week	Mid-Week	Last Week
BIM	U	3211971.5	2293014.5	2411750.5
	P	5e-148***	5e-250***	5e-259***
BD	U	476093	387020	561591
	P	3e-103***	6e-225***	9e-209***
SC	U	21763.5	13247.5	16621.0
	P	4e-23***	1e-52***	2e-64***

SH	U	3304763.0	1772883.0	1437034.5
	P	1e-97***	6e-286***	0***
TMF	U	2160133.0	1436888.0	1370678.0
	P	1e-85***	2e-182***	6e-264***

*: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$.

5.4.1.2. Attempts

In Table 5.5, the statistical analysis results of comparing the number of attempts (either correct or wrong) for non-paying learners and certificate purchasers are provided. As in the earlier results in 5.4.1.1, bold values refer to the most significant value in each scenario (week) in each course, whereas italic denotes the *p* value. The statistical significance has also followed a similar trend to the number of accesses, where course purchasers had larger statistics (number of attempts) towards the end of the course. All the statistical analysis results were significant, with *p* value ranging between 9e-38 and zero for SC 1st week and the last week of BIM, SH and TMF, respectively.

Table 5.5. Comparison of the number of attempts for non-paying learners and certificate purchasers at three different time points of the course.

Course	**Measure**	**1st Week**	**Mid-Week**	**Last Week**
BIM	U	4080976.0	2682326.0	3333306.5
	P	1e-117***	0***	0***
BD	U	545904	555967	747421
	P	4e-136***	5e-269***	2e-279***
SC	U	23504.5	13759.0	19366.5
	P	9e-38***	2e-77***	1e-88***
SH	U	3828157.5	1844641.5	1984241.0
	P	8e-84***	0***	0***
TMF	U	1931024.0	1367316.5	1855830.0
	P	4e-130***	4e-257***	0***

*: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$.

5.4.1.3. Correct Answers

Table 5.6 shows the statistical analysis results of comparing the correct answers for non-paying learners and certificate purchasers at three different time points of the courses analysed. Regarding statistical significance, correct answers also followed a similar trend to the number of accesses and attempts, where course purchasers had larger statistics (number of correct answers) towards the end of the course. As shown in Table 5.7, all the statistical analysis results were significant, with *p* values ranging between 4e-42 and zero for the SC 1st week and the last week of BIM and SH, respectively.

Table 5.6. Comparison of the number of correct answers for non-paying learners and certificate purchasers at three different time points of the course.

Course	Measure	1st Week	Mid-Week	Last Week
BIM	U	3676853.0	2570707.5	3264674.5
	P	3e-155***	0***	0***
BD	U	531926	539949	752898
	P	7e-142***	7e-277***	9e-279***
SC	U	20705.5	13695.0	19429.5
	P	4e-42***	1e-77***	1e-88***
SH	U	2879619.5	1692500.0	1851654.5
	P	5e-140***	0***	0***
TMF	U	2228499.5	1430002.5	1916800.0
	P	8e-116***	5e-252***	1e-308***

**: P < 0.05, **: P < 0.01, ***: P < 0.001.*

5.4.1.4. Wrong Answers

Wrong answers are the fourth primary variable adopted for predicting certification in the present experiment. Table 5.7 shows the statistical analysis results comparing the number of wrong answers for non-paying learners and certificate purchasers at three different time points of the courses analysed. In terms of statistical significance, wrong answers also followed a similar trend to the number of accesses, attempts and correct answers, where course purchasers had larger statistics (number of wrong answers) towards the end of the course. As shown in Table 5.8, all the statistical analysis results were significant, with *p* value ranging between 2e-11 and zero for SC 1st week and SH last week, respectively.

Table 5.7. Comparison of the number of wrong answers for non-paying learners and certificate purchasers at three different time points of the course.

Course	Measure	1st Week	Mid-Week	Last Week
BIM	U	4532257.5	3621172.0	5021527.0
	P	1e-95***	9e-224***	8e-256***
BD	U	1125357.5	1292662	863940
	P	8e-67***	1e-95***	1e-250***
SC	U	59297.0	21375.5	23286.5
	P	2e-11***	1e-69***	5e-83***
SH	U	4198990.5	2045237.0	2292388.5
	P	3e-68***	5e-265***	0***
TMF	U	2082685.5	1540113.5	2005657.5
	P	5e-125***	6e-249***	4e-310***

**: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$.*

5.4.2. Prediction Performance

The results shown in Table 5.8 compare the performance of our selected classifiers based on Rec and BA performance metrics. These results answer our second research question, on whether the learner clickstreams can be used to predict paid certification in MOOCs. The results achieved a promising BA across the five domain-varying courses, ranging from 0.77 to 0.95. The classifiers performed differently based on the course analysed, where SVC performed the best in BIM, ET in BD and SP, and LR in CS and TMF. In general, the improvement in the performance of the classifiers was lower towards the end of the courses compared to the difference between the first week only and the first half of the course. This may indicate that course purchasers exert more effort until they reach the minimum requirements for certification (typically just after the middle of the course). After that, the level of interest in the course content in terms of access, question answering, and time spent learning is reduced; hence, at this stage, activities that are more similar to non-paying learners are performed even by the paying course takers.

Table 5.8. clickstream-based Learner classification results distributed by course, class 0 = non-paying learners, class 1 = paid learners.

Course	Classifier	1st Week			Mid-Week			Last Week		
		Rec_0	**Rec_1**	**BA**	**Rec_0**	**Rec_1**	**BA**	**Rec_0**	**Rec_1**	**BA**
BIM	RF	0.61	0.95	**0.78**	0.79	0.85	0.82	0.80	0.85	0.83
	ET	0.60	0.95	0.77	0.80	0.82	0.81	0.81	0.82	0.81
	LR	0.60	0.95	**0.78**	0.78	0.86	0.82	0.80	0.86	0.83
	SVC	0.59	0.96	**0.78**	0.79	0.87	**0.83**	0.80	0.87	**0.84**
BD	RF	0.78	0.96	**0.87**	0.87	0.86	0.86	0.87	0.95	**0.91**
	ET	0.76	0.98	**0.87**	0.85	0.90	**0.88**	0.86	0.95	**0.91**
	LR	0.76	0.98	**0.87**	0.86	0.88	0.87	0.86	0.95	**0.91**
	SVC	0.76	0.98	**0.87**	0.85	0.90	0.87	0.85	0.95	0.90
CS	RF	0.78	1.00	**0.89**	0.90	0.90	**0.90**	0.90	1.00	**0.95**
	ET	0.78	1.00	**0.89**	0.89	0.90	0.89	0.89	1.00	**0.95**
	LR	0.78	1.00	**0.89**	0.90	0.90	**0.90**	0.90	1.00	**0.95**
	SVC	0.78	1.00	**0.89**	0.90	0.85	0.87	0.89	1.00	**0.95**
SP	RF	0.55	0.98	**0.77**	0.79	0.96	0.87	0.84	0.91	0.87
	ET	0.55	0.98	**0.77**	0.79	0.96	**0.88**	0.84	0.92	**0.88**
	LR	0.58	0.95	0.76	0.84	0.90	0.87	0.84	0.90	0.87
	SVC	0.55	0.98	**0.77**	0.79	0.96	0.87	0.84	0.91	0.87
TMF	RF	0.66	0.96	0.81	0.80	0.93	**0.86**	0.85	0.86	**0.86**
	ET	0.66	0.98	**0.82**	0.81	0.89	0.85	0.84	0.86	0.85
	LR	0.66	0.98	**0.82**	0.80	0.93	**0.86**	0.84	0.86	0.85
	SVC	0.66	0.98	**0.82**	0.81	0.89	0.85	0.84	0.86	0.85

The BA metrics on the “first week-only” data were relatively promising (0.82 on average) for a very early prediction, compared to the first half of the week and the whole course data (0.86 and 0.88, respectively). However, at the class level, the average of Rec_0 (non-paying learners) in the first-week scenario was 0.67 compared with 0.83, using data from the first half of the course. We thus believe that mid-week should be considered a more reliable point of the course for classifying learners and consequently providing any planned intervention. In class weighing, we have used both approaches of automatic balancing by the algorithm (balanced) and manually assigning a weight for each class until the balanced optimal Recall is reached for both classes. This significantly affects the learning process; for instance, SC was the most challenging course to balance learning, as testing was done on a small set of four learners only.

As discussed in Chapter 3, the correlation between the time spent on the course content and certification was statistically analysed by various previous works, including those by Cobos and Jurado (2018), who used the learners' time spent on assignments and videos and Goli, Chintagunta and Sriram (2019) who used the total time spent (minutes) and the average session duration (minutes). Tian et al. (2017) analysed the correlation between the time learners spend on content in days and certification, whereas Qiu et al. (2016) examined the impact of effective learning time spent and certification attainment. However, the time spent on content as a feature for building a paid certification predictive model has not been used on MOOC data on a large scale. This highlights the contribution of the present experiment, which computed this feature of relatively large data of 23 runs of 5 courses.

As there is no predefined strategy for selecting the most suitable algorithms for a specific classification task, this study followed an adequately justified approach to selecting algorithms. For instance, this experiment was based on using the data as-is, without oversampling any of the training dataset classes. As a result, the classification algorithms adopted in this experiment included class weighting among their fine-tuning parameters. The *class_weight* parameter uses the number of instances in each class to adjust weights inversely proportionally to each class size. Both automatic adjustment (balanced) and manual (giving each class a weight percentage) were assigned with iterative runs to reach optimal results. Other fine-tuned parameters included regularisation control (C, where smaller values specify stronger regularisation) and kernel in LR and SVC classifiers. For tree-based classifiers, the number of trees in the forest (n_estimators) and the maximum depth of the tree (max_depth) were tuned in RF and ET classifiers. These parameters significantly impacted the learning process in each of the corresponding classifiers mentioned above and were iteratively assigned integer, float, or Boolean values until an optimal model performance was reached.

Regarding generalisability, while the dataset we used was obtained from one platform only, it spans various runs (23) of 5 different MOOCs, covering four distinct disciplines (literature, psychology, computer science, and business). This allowed us to longitudinally observe the changes in learners' activities from a paid certification perspective, as further discussed in Section 5.3.1 and predict paid certification based on a rich source of data. As shown in Table 3.2, most of the surveyed works were based either on one course or run/iteration. Another novelty of the present thesis is predicting paid certification in MOOCs at an early stage (starting with using data from the first week of the course only). Our analysis showed that only ten studies adopted early prediction, out of which only one study explicitly predicted paid certification at an early point of the course using data from one run of an edX course. However, while edX has been extensively analysed (17 out of 25 studies), FutureLearn has never been explored from a paid prediction perspective, although it is considered the largest MOOC platform launched outside the United States.

5.5. Epilogue

This chapter focuses on using learners' weekly activities (accesses, attempts, and correct and wrong answers) and the time spent on each learning step to temporarily predict paid certification in MOOCs. The experiment is based on 23 runs of 5 MOOCs obtained from the relatively unexplored MOOC platform of FutureLearn to first statistically measure the difference in terms of activities between non-paying learners and course purchasers and then predict course purchasability using various ML classifiers. In the next chapter, a MOOC sentiment classifier using reviews and comments data from various MOOC platforms is introduced, in addition to a comparison of the performance of different lexicon and DL-based classifiers, as a step towards better predicting certification purchase.

Chapter 6 : MOOCSent: A Sentiment Predictor for Massive Open Online Courses

6.1. Prologue

After conducting an initial analysis involving straightforward predictions based on a limited set of clickstream variables in Chapter 5, our focus shifts in the current chapter, towards exploring a more complex predictor (i.e., discussion forums) of paid certification in MOOCs. However, since Sentiment Analysis (SA) has been identified as a key determinant of learners' success in MOOCs (Kastrati *et al.*, 2021; Dalipi, Zdravkova and Ahlgren, 2021), MOOCSent was built (as discussed in the present chapter) before delving into the examination of learners' discussion forums. Since the forum discussion dataset used in this thesis lacked labelled sentiments of learners, therefore, as a preliminary step towards harnessing MOOC discussion forums for certification prediction in Chapter 7, assigning sentiments to learners' posts, encompassing both comments and replies, was first conducted using MOOCSent. This sentiment labelling process aimed to enhance the richness of our training dataset and subsequently improve the predictive model's performance.

Thus, this chapter introduces a cross-platform MOOCs sentiment classifier, using over 1.2 million human-annotated learners’ comments and reviews obtained from 633 MOOCs. The classifiers adopted in this experiment varies from lexicon and rule-based (LRB) tools, to the more advanced language model of BERT.

The purpose of constructing the MOOCSent model was to use the labelled dataset of 1.2 million MOOC reviews to train the model and subsequently apply it to label textual data from FutureLearn's learners (comprising comments and replies) with estimated sentiments. These estimated sentiments serve as input features in the third experiment in Chapter 7, which focuses on predicting certification outcomes based on discussion forums.

6.2. Introduction

The terms “Sentiment Analysis” (SA) and “Opinion Mining” are used interchangeably (Medhat, Hassan and Korashy, 2014), together with other terms with the same principal aim (Bonta and Janardhan, 2019). They are defined as the process of computational evaluation and classification of opinions from unstructured text to determine if they tend towards positive, negative, or neutral sentiments (Ahuja and Dubey, 2017). Sentiment analysis has become a valuable tool in solving a wide range of problems by extracting opinions and making decisions across different disciplines and fields, including sociology, marketing and advertising, psychology, economics, political science, and others (Hutto and Gilbert, 2014). Its widespread use can be attributed to the fact that opinions are important factors affecting human behaviours (Zhang, Wang and Liu, 2018). Adopting SA, particularly in the education domain, is an essential but a complicated task due to the specific nature of textual data and the volume of information learners generate on online learning platforms (Kastrati *et al.*, 2021). Additionally, annotating learners’ expressed opinions and sentiments in MOOCs is a time-consuming and labour-intensive task. This might not be the case with small-scale MOOCs which typically contain relatively small number of textual inputs. However, manual labelling seems impractical for many online courses and specifically for MOOCs (Kastrati, Imran and Kurti, 2020).

Another challenge in the SA domain is the struggle to identify sarcasm and irony in text, where, in some cases, the intended sentiment may be exactly the literal contrast to the words posted. For example, sarcastic sentences such as "Great job!" in a negative context might be misinterpreted if the classifier cannot capture the subtle nuances of language (Ilavarasan, 2020). Another challenge lies in dealing with context-dependent expressions and ambiguous language. While sentiment lexicons play an essential role in sentiment detection tasks by providing sentiment information for words, the sentimental ambiguity - with these lexicons having one sentiment polarity for each word - is typically ignored. Thus, incorporating POS chunks with the words is expected to solve the ambiguity of lexical sentiments where POS of context can be used for calculating the sentiment (Yin *et al.*, 2020)

Cultural and linguistic variations are also among the main challenges and can further complicate sentiment analysis since the same word/phrase may convey different sentiments in different cultural contexts (Mirza, Lukosch and Lukosch, 2023). This is related to cross-cultural polarity measurement using emotion detection, where sentiment analysis techniques are employed to measure the sentimental tone in multicultural or multilingual texts. One way to address this challenge is to compare and contrast the emotions expressed in different cultures or identify cultural-specific sentiment patterns using deep learning approaches (Mirza, Lukosch and Lukosch, 2023). These challenges underscore the need for advanced techniques and data-rich models to enhance the accuracy and robustness of sentiment analysis in real-world applications.

With regard to the approaches used for SA, they include lexicon-based (rule-based), conventional, and deep ML models and, recently, complex language models such as BERT. While current and common lexicon-based classifiers (such as VADER (Hutto and Gilbert, 2014) and TextBlob (Loria, 2018)) and conventional ML models such as NB) have been the most commonly used classifiers in the education domain (Dalipi, Zdravkova and Ahlgren, 2021), it is not obvious which is more appropriate and whether recent language models such as BERT can outperform these approaches.

There are various concerns regarding the effectiveness and generalisability of the current literature on SA in MOOCs. This includes (1) using one single source of data (previous literature on sentiment classification in MOOCs was based on single platforms only, Coursera and edX being the most popular platforms (Bulusu and Rao, 2021)) and hence less generalisable with relatively low number of instances compared to our obtained dataset; (2) lower model outputs, where the majority of the surveyed models in Section 6.3 are based on 2-polar classifier (positive or negative); (3) disregarding important sentiment indicators, such as emojis and emoticons, during text embedding; and (4) reporting average performance metrics only, preventing the evaluation of model performance at the level of class (sentiment).

In this chapter, we further fill this gap by comparing the various and currently widely used NLP methods (TextBlob, VADER, Stanza, and NB) for SA to validate these tools in the educational sector, especially in discussion forums in MOOC platforms. In addition, we propose MOOCSent using the

BERT-based model to predict sentiment in a massive dataset of around 1.2 million learners' comments so as to find the most suitable model for sentiment prediction. Thus, the research question this chapter tackles is as follows:

- *RQ3: Can course reviews obtained from multiple MOOC platforms be used to build a reliable sentiment classifier?*

6.3. Related Work

The researchers' interest in SA began in the early 1990s (Ahuja and Dubey, 2017). Later, in 2000, it become one of the most active areas in NLP (Bonta and Janardhan, 2019). It has been employed in numerous studies of educational data mining using NLP methods. Bakharia (2016) used 3 cross-domain MOOCs (education medicine and technology) to develop a 2-polarity learner sentiment classifier (negative or positive). In this study, several ML algorithms such as NB, SVM, and RF have been used. The data source was Stanford MOOCPosts dataset which contains approximately 30,000 forum posts. Although all the classifications achieved an accuracy of over 0.70, only the average classification accuracy has been reported. In the transfer learning research, Wei *et al.* (2017) investigated cross-domain classification using deep neural network techniques based on CNN and LSTM to determine the polarity of the sentiment for a highly imbalanced dataset (17,936 (85%) positive posts and 3,157 (15%) negative posts). The reported metric was the overall accuracy of the different models, and the best preforming one achieved an overall Acc. of over 0.85.

Clavié and Gal (2019) built two models, namely EduBERT and EduDistilBERT, to classify confusion, sentiment, and urgency. The dataset used for these models was derived from 11 Stanford courses and 18 courses from various public universities in the UK and the US. The best-performing algorithm was EduBERT for sentiment classification, achieving 89.78%. However, the study did not present the performance of the models for each class. Chen *et al.* (2019) used three randomly selected courses from the Stanford dataset mentioned above to predict SA. Three experiments were conducted to verify the effectiveness of the proposed model (1) using traditional supervised methods, that is, Random Forest and SVM (RBF); (2) using CNN with the pre-trained word vectors trained on GN and ELMo; and (3) using a semi-supervised learning model with 30% labelled data. The latter method improved the model accuracy and the F1 score by 2.8% and 3.2%, respectively.

Kastrati *et al.* (2020) built an aspect-based opinion and sentiment model for mining learners' comments to predict their attitudes of these commented aspects. The dataset used for implementing the model contained over 21,000 manually annotated Coursera learners' reviews. Various conventional and deep models including DT, N, SVM, GB, and CNN were adopted in this study. The experiment was replicated (Kastrati, Imran and Kurti, 2020) but with a larger dataset of 111,000 sentiment-annotated reviews from Coursera and other traditional classes. The purpose was to use this classifier to label other unlabelled learners' reviews, consequently reducing the need for labour-intensive manual annotation. The proposed model achieved a promising performance in identifying aspect categories and sentiments within learners' reviews. Yet, adapting contextualised word-embedding models such as BERT is in the pipeline to further improve the performance of the model.

Onan (2021) used a crawled CourseTalk dataset of 93,000 course reviews with several conventional, ensemble, and deep classifiers to build a sentiment classification model. The empirical analysis indicates that deep models outperform the other architecture with LSTM being the best sentiment classifier. Moreno-Marcos *et al.* (2018a) compared the performance of LR, SVM, DT, RF, NB, and SentiWordNet in detecting learners' sentiments. The experiment used around 13,000 comments obtained from an edX course and found RF to be the best classifier. The analysis found that positivity correlates with different time points of the course; for instance, learners tend to post more positive comments at the beginning of the course, whereas more negative posts were observed at critical stages of the course such as the deadlines for peer-review assessments.

In the MOOCs domain there are some efforts of using SA on forum content for various purposes. For example, a popular target is using sentiment as a feature to predict learner attrition in MOOCs (Chaplot, Rhim and Kim, 2015). Literature shows a correlation between learners' sentiment and their performance in MOOCs such as course quizzes, homework assignment and course completion. Wen, Yang and Rose (2014) examined the correlation between learners' sentiments extracted from the comments posted weekly and the weekly dropout rate in 3 MOOCs. The study found that correlation is significant, recommending further consideration of learners' weekly textual interaction. Similarly, Tucker, Pursel and Divinsky (2014) used a sentiment analyser (the Semantic Orientation CALculator (SO-CAL)) to mine a MOOC discussion forum. Next, quantified the learners' sentiment impact on learner performance (grades, assignment, quizzes, exams) and learning outcomes. The purpose of this study was to determine whether there is a correlation between textual content expressed in the discussion forum and learner performance and learning outcomes. The results showed that learner positive sentiment slightly correlated with performance in quizzes, whereas a stronger correlation was found between negative sentiment and homework assignment. Table 6.1 shows a list of the previous works compared to our SA predictor, MOOCSent.

Table 6.1. Sentiment prediction models vs MOOCSent

Cite.	Dataset	Source	#Courses	Classifiers	Metrics
(Moreno-Marcos *et al.*, 2018a)	≈13k	edX	1	LR, SVM, DT, RF, NB, SentiWordNet	AUC, Kappa
(Bakharia, 2016)	≈18k	Stanford	3	NB, SVM, RF	Acc
(Wei et al., 2017)	≈18k	Stanford	3	CNN, LSTM	Acc
(Li *et al.*, 2019b)	≈19k	China MOOC	n/a	Lexicon (DUTIR), CNN, GRU, BERT	F1, Acc.
(Clavié and Gal, 2019)	n/a	Stanford & other	29	BERT	Acc.
(Kastrati *et al.*, 2020)	≈21k	Coursera	n/a	DT, NB, SVM, GB, CNN	Prec., Rec., F1
(Lundqvist, Liyanagunawardena and Starkey, 2020)	=25k	FutureLearn	1	VADER, Wilcox Pratt	P
(Tucker, Pursel and Divinsky, 2014)	≈26k	Coursera	1	SO-CAL	Corr.
(Chen *et al.*, 2019)	≈30k	Stanford	11	RF, SVM, CNN	F1, Acc.
(Wen, Yang and Rose, 2014)	≈36k	Coursera	3	Survival Analysis	p
(Munigadiapa and Adilakshmi, 2022)	≈80k	Stanford, IMDB	11	NB, SVM, LR, CNN, LSTM	Prec., Rec., F1, Acc.
(Onan, 2021)	=93k	CourseTalk	n/a	KNN, RF, NB, AdaB, SVM, CNN, RNN, LSTM, GRU	F1, Acc.
(Kastrati, Imran and Kurti, 2020)	≈111k	Coursera & other	n/a	CNN, LSTM	Prec., Rec., F1
(Yang, 2021)	=150k	icourse163	n/a	GRU	Prec., Rec., F1
MOOCSent	**≈ 1.2m**	**Coursera, Udemy, FutureLearn, Stanford**	**633**	**TextBlob, VADER, CNN (Stanza), NB, BERT**	**Prec., Rec., F1, Acc.**

6.4. Methodology

6.4.1. Data Collection

Here, we propose a cross-platform MOOCs sentiment classifier using almost 1.2 million human-annotated learners' comments obtained from 633 MOOCs delivered via Coursera, Udemy, FutureLearn and Stanford University platforms. This makes our dataset the largest MOOC datasets collected for SA.

6.4.1.1. Coursera, Udemy and FutureLearn (training data)

The reviews dataset contains 1,250,830 reviews scraped from 622 Coursera, Udemy, and FutureLearn courses, in addition to the learner rating. The distribution of instances per class is shown in Table 6.2. After completing a given course, the learners were asked to provide their review about the course together with a 3-point-Likert-scale rating of the learner's sentiment towards the course (positive, neutral, or positive). The rating makes the dataset already sentiment annotated, thus saving a great deal of time that would otherwise be spent on manual annotation.

6.4.1.2. Stanford university (test data)

This dataset, dubbed the Stanford MOOCPosts Dataset, is available for academic researchers by request. It contains anonymised learners' posts in English from the discussion forums of 11 Stanford University online courses spanning over 3 different domain areas: education, humanities/sciences, and medicine (Agrawal et al., 2015). These textual posts were labelled by 3 human annotators across 6 dimensions (Opinion, Question, Answer, Sentiment, Confusion, and Urgency). For sentiment, the range was from 1 to 7, with 1 = negative, 7 = positive, and 4 = neutral. For a better and fairer comparison with other approaches, we simplified by converting the 7-point scale into 3 classes as: Negative → sentiment (1–7) < 4, Neutral → sentiment (1–7) $\in$ [4, 5), Positive → otherwise.

Therefore, the distribution of the classes is as follows: 4,387 instances in the negative class, 20,557 in the neutral class, and 4,653 in the positive class (Table 2).

Table 6.2. Statistics of the experiment datasets

Dataset	**#Negative**	**#Neutral**	**#Positive**	**Total**
Coursera	29,234	32,073	706,966	768,273
Udemy	17,036	38,349	304,172	359,557
FutureLearn	1,259	3,853	117,888	123,000
Stanford	4,387	20,557	4,653	29,597
Total	51,916	94,832	1,133,679	1,280,427

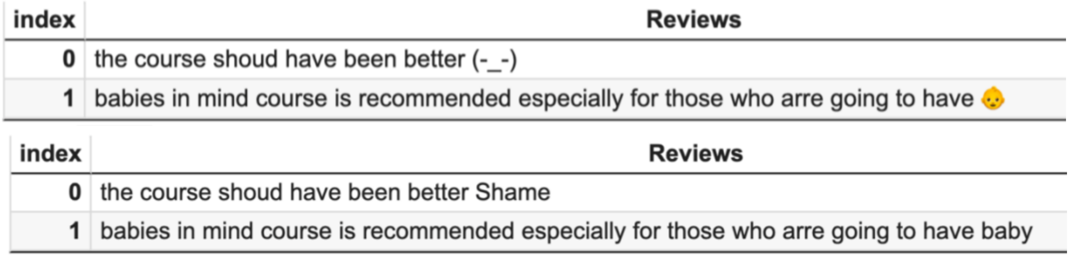
6.4.2. Data Preprocessing

As shown in Table 6.2, the number of instances of the negative and neutral classes are relatively very low. Thus, augmenting the training data is a crucial step to help influence the model performance and prevent model overfitting. There are various data augmentation techniques such as random word insertion, random word deletion, and back translation. However, these methods may introduce semantic (meaning) variance and hence misrepresentation of learners' reviews (Shorten, Khoshgoftaar and Furht, 2021). Therefore, we adopted one of the most used semantically preserving augmentation method called WordNet, which augments text by replacing words with their synonyms of the same POS from the WordNet thesaurus (Miller et al., 1990; Bayer, Kaufhold and Reuter, 2022).

Data preprocessing also included removing unwanted characters, such as HTML/XML, punctuations, and non-alphabets, by using regular expressions. The last step contained removing stop-words, lowering the cases of characters, reforming contractions into the original words, and grammar correction.

Some preprocessing steps such as stemming, lemmatisation, and stopwords removal were skipped for BERT. These steps are useful while using lexicon-based predictors as the weighting scheme is usually *tf-idf-* or countervectoriser-based. Thus, the text contextuality is not important. However, context-aware semantic models such as BERT process stopwords-included context like the negation words (e.g. never, nor, not) to capture knowledge and can thus learn the true learner's sentiment. Therefore, stopwords receive as much attention as the rest of the textual inputs by BERT due to its important role in structuring semantic representation (Qiao et al., 2019).

Emojis and emoticons play a significant role in analysing sentiment; nevertheless, they are not by default recognised by language models. This includes advanced NLP models such as BERT, where emojies and emoticons are tokenised as unknown [UNK]. Therefore, we use UNICODE_EMOJI and EMOTICONS_EMO lexicons to address this issue. These lexicons contain 221 emoticons and 3,521 emojis with their corresponding explanatory words/phrases, which were used to convert emojis and emoticons into tokenisable inputs. Figure 6.1. depicts an example of converted emoji and emoticon into corresponding explanatory words.



index	Reviews
0	the course shoud have been better (-_-)
1	babies in mind course is recommended especially for those who arre going to have 👶

index	Reviews
0	the course shoud have been better Shame
1	babies in mind course is recommended especially for those who arre going to have baby

Figure 6.1. An example of converting emojis and emoticons into corresponding explanatory words.

6.4.2.1. Dealing with Bias

Addressing bias in text-based tasks is a sophisticated and ongoing challenge. This includes finding actionable techniques to promote fairness and ensure that the developed models are equitable and provide reliable results. It also includes improving the data collection methods for promoting transparency in NLP model development. There are some occasions where bias in NLP tasks is expected. This includes but is not limited to (1) using imbalanced sets of the different classes of the training dataset, (2) passing improperly tokenised texts, and (3) human annotating the model inputs, which increases the chance of unintentional encoding of annotators' biases (Baker and Hawn, 2021).

The current experimental methodology followed various techniques for mitigating potential bias. For instance, the data collected for building the predictive model was obtained from different courses (633) of different sources (4 platforms), making it not only the largest dataset collected for SA in the field of MOOC analytics but also the most diverse in terms of course types and disciplines (see table 6.1 for more comparison). This helped build an SA classifier trained on four of the largest MOOC platforms (Coursera, Udemy, FutureLearn, Stanford), achieving a higher level of generalisability and reducing the model risk of being overfitting on a single platform.

Another procedure followed to reduce potential bias is balancing the training data before training the model. As shown in Table 6.2, the number of instances of the negative and neutral classes is relatively low, marking only 12% combined, whereas positive comments mark the remaining portion.

Thus, augmenting the training data (namely negative and neutral classes) is crucial to help influence the model performance and prevent any potential bias. There are various data augmentation techniques, such as random word insertion, random word deletion, and back translation. However, these methods may introduce semantic (meaning) variance and consequently misrepresent learners' reviews (Shorten, Khoshgoftaar and Furht, 2021). Therefore, we adopted one of the most used semantically preserving augmentation methods called WordNet, which augments text by replacing words with their synonyms of the same POS from the WordNet thesaurus (Miller et al., 1990; Bayer, Kaufhold and Reuter, 2022).

Having our collected dataset labelled by the learners themselves is another factor for mitigating any potential bias. Human manual labelling for training may introduce bias because annotators may unintentionally encode their biases. Since learners themselves annotate the data adopted, this typically represents higher accuracy in terms of annotation (Malko *et al.*, 2021) and promotes the fairness of the results. Also, key sentiment indicators (i.e., emojis and emoticons) were properly encoded before tokenisation. Although emojis and emoticons play a significant role in analysing sentiment, they are not, by default, recognised by language models. This includes advanced NLP models such as BERT, where emojis and emoticons are tokenised as unknown [UNK]. Therefore, we use UNICODE_EMOJI and EMOTICONS_EMO lexicons to address this issue. These lexicons contain 221 emoticons and 3,521 emojis with their corresponding explanatory words/phrases, which were used to convert emojis and emoticons into tokenisable inputs. Figure 6.1. depicts an example of converted emoji and emoticon into corresponding explanatory words. This can help the model train on the entirely representative texts typed by learners instead of assigning random tokens to these key inputs and, therefore, become more reliable and fairer.

6.4.3. Sentiment Classification Methods

6.4.3.1. TextBlob

TextBlob is an open-source text-processing Python library that allows one to perform several tasks, including noun phrase extraction, translation, part-of-speech tagging, sentiment analysis, tokenisation, and spelling correction. TextBlob is part of the well-known natural language toolkit (NLTK) and helps in reducing the computational cost of analysis. The tool generates a float value of a confidence level (between -1 and 1) for each text inserted and later annotates it as positive if > 0, negative if < 0, or neutral if = 0. These default thresholds however can be manually adjusted.

TextBlob assesses sentiment via returning a tuple of form (polarity, subjectivity, and assessments), where polarity and subjectivity float within a range of -1 and 1, with 0 being very objective and 1 being very subjective; assessments is a list of polarity and subjectivity scores for the assessed tokens.

6.4.3.2. VADER

Valence Aware Dictionary and Sentiment Reasoner (VADER) is a social media-based tool for general sentiment analysis. This open-source lexicon and rule-based tool uses a mix of qualitative and quantitative methods (a gold-standard list of lexical features along with their associated sentiment intensity measures), which are specifically attuned to sentiment in microblog-like contexts. Afterwards, the lexical features are combined, with consideration of five general rules, which embody grammatical and syntactical conventions, to express and emphasise sentiment intensity (Hutto and Gilbert, 2014). Similar to TextBlob, VADER generates a sentiment confidence level for each analysed text and allows resetting the thresholds of < 0, = 0, and > 0.

6.4.3.3. Stanza

Stanza is also an open-source Python natural language processing toolkit which can be used for lemmatisation, tokenisation, part-of-speech, multi-word token expansion, morphological feature tagging, sentiment tagging, dependency parsing, and named-entity recognition. This toolkit uses CNN for its architecture and massively supports more than 60 human languages. It was trained on 112 datasets, including the Universal Dependencies treebanks and other multilingual corpora. In comparison with the lexicon and rule-based tools, Stanza features a language-agnostic fully neural pipeline for text analysis, including a native Python interface to the widely used Java Stanford CoreNLP software. This makes it capable of more functionality and more advanced tasks, like relation extraction and co-reference resolution (Qi et al., 2020).

6.4.3.4. Naïve Bayes (NB)

Naïve Bayes (or simple independent Bayes) is a Bayes' theorem-based probabilistic Bayesian classification technique. NB classifier assumes strong (naïve) independence between the features (i.e. the existence of one feature in a category has no relevance to the existence of other features). Regardless of its simplicity among other Bayesian network models, NB shows good performance in some sophisticated tasks such as multi-class and text classification tasks (Rish, 2001).

6.4.3.5. BERT

BBERT is one of the most advanced language representation models for a broad range of NLP tasks, such as question answering, language inference, and sentiment analysis. BERT is developed via pre-training a deep bidirectional representation by jointly conditioning a two-way context for all layers. BERT has two parameter-intensive settings: (1) BERTBASE: 12 layers, 768 hidden dimensions, and 12 bidirectional self-attention heads with 110 million parameters and (2) BERTLARGE: 24 layers, 1,024 hidden dimensions, and 16 bidirectional self-attention heads (in transformer) with 350 million parameters. BERT is trained from unlabelled data obtained from Wikipedia (2,500M words) and BookCorpus (800M words) (Devlin et al., 2018).

6.4.3.5.1. Embedding Layer

BERT, in contrast to traditional embedding methods of Word2Vec or GloVe, provides a multiple, context-independent representation for each token. Its embedding layer takes a learner's comment as input and calculates the token-level representations via the extracted knowledge of each sentence from the entire comment (Li *et al.*, 2019a). Firstly, we pack the input features as follows:

$$E_0 = (e_1 \dots e_n) \tag{6.1}$$

Where, e_n ($n \in$ [1,N]) is the combination of the token embedding, position embedding, and segment embedding corresponding to the input token Xn. Note that [CLS] is a special symbol embedded prior to each comment input, and [SEP] is a special separator token splitting each comment/review into several sentences. Text input is tokenised and then converted to token IDs. Figure 6.2 shows a snapshot of the first layer of representation (token embedding), which include a tensor of token IDs (numerical tokens plus BERT special tokens such as CLSs and SEPs).

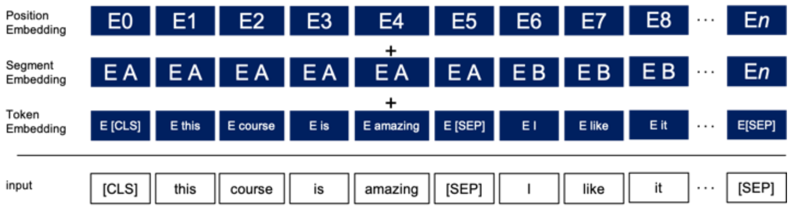
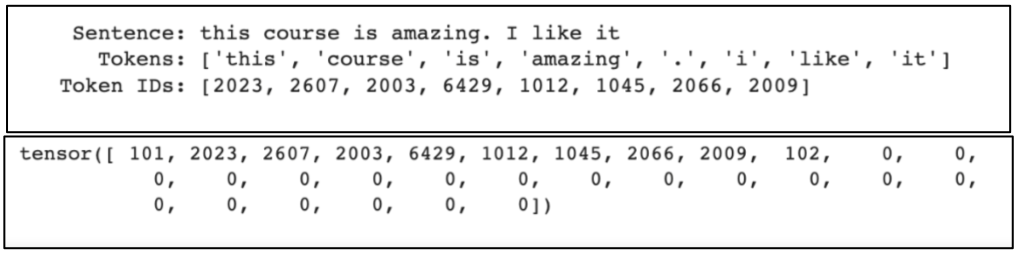


Figure 6.2. BERT input representation (The sum of the three embeddings)

This tensor is padded into a length of max_len, 30, in the example shown in Figure 6.3.



```
 Sentence: this course is amazing. I like it
   Tokens: ['this', 'course', 'is', 'amazing', '.', 'i', 'like', 'it']
Token IDs: [2023, 2607, 2003, 6429, 1012, 1045, 2066, 2009]
```

```
tensor([ 101, 2023, 2607, 2003, 6429, 1012, 1045, 2066, 2009,  102,    0,    0,
           0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,    0,
           0,    0,    0,    0,    0,    0])
```

Figure 6.3. BERT-based text tokenisation

The next step corresponds to the L transformer layers, where the token-level features are refined, layer by layer. Specifically, the representations

$$H_1 = (h_{11} \dots h_{1t}) \tag{6.2}$$

at the *l*-th (*l* ∈ [1,L]) layer which are calculated as below:

$$H_i = Trm_i\ (h_i - 1) \tag{6.3}$$

Where Hl is the contextualised representation of the input tokens used for performing the predictions. The last_hidden_state is a sequence of hidden states of the last layer of the model. BertPooler is applied on the last_hidden_state to obtain the pooled_output.

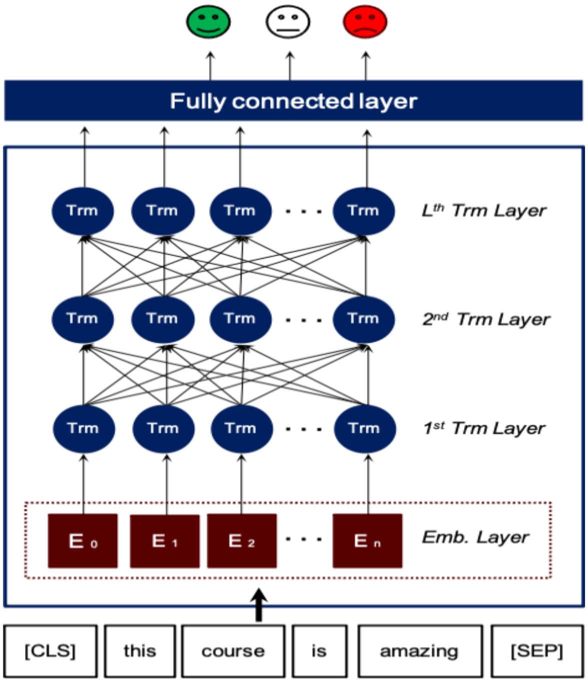


Figure 6.4. BERT-based sentiment prediction model.

6.4.3.5.2. *Fine tuning*

We ran several experiments with different parameters, namely the type of BERT (Large Cased, Large Uncased, Base Uncased, Base Cased), maximum sequences length (between 100 and 256 sequences), Adam learning rate (ranging from 2e-5 to 5e-5), batch size (from 8–32), and the number of Epochs (between 2 and 5). We use the pre-trained uncased BERT-base model for fine-tuning. Taking into consideration the computational cost of BERT as a complex, large model along with the recommended parameters by the model authors, we set the above parameters as follows:

- Early_stopping: To avoid overfitting, an early stopping threshold was specified for when the training accuracy reaches 0.95.
- Training model = BERT Base Uncased.
- Max_len = 200, based on the distribution of sequence lengths (see Figure 5).
- #Epoch = 2, in association with the early stopping threshold specified earlier.
- #Transformer layers = 12, with 768 hidden dimensions, 12 bidirectional self-attention heads.

- Batch_size = 16.
- Learning_rate = 2e-5.

As BERT works with fixed-length sequences, we set the *max_len* = 200 based on the token length of each review as illustrated in Figure 6.5. The running time of various experiments ranged from 7 h, 10 min, and 43 s up to 16 h, 8 min, and 35s based on the parameters specified. We used Tesla V100-SXM2 32GB GPU to run our experiments.

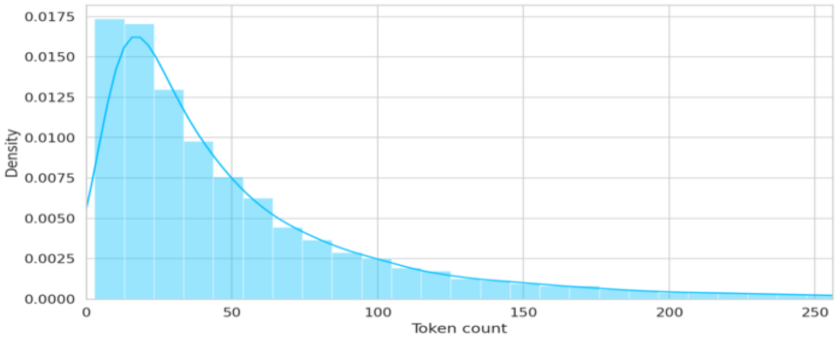


Figure 6.5. Distribution of sequence lengths (tokens).

6.5. Results

Table 6.3 shows the results of our sentiment prediction model using TextBlob, VADER, Stanza, NB, and BERT. The negative, neutral, and positive metrics denote the recall for each class, whereas WF1 denotes the weighted F1. Furthermore, macro F1 and accuracy (Acc.) were also reported, to explore the overall performance of the model. F1 can be considered a trade-off between precision and recall. BA is the average of recall obtained in each class, which equals the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate). Acc. calculates the model's overall performance, giving the same weighting for all classes by dividing the total number of correctly predicted instances by the total number of instances.

Table 6.3. Sentiment prediction results using TextBlob, VADER, Stanza, NB, and BERT.

Model	**Negative**	**Neutral**	**Positive**	**BA**	**WF1**	**F1**	**Acc**
TexBlob	0.74	0.65	0.78	0.72	0.71	0.68	0.70
VADER	0.63	0.83	0.78	0.74	0.78	0.75	0.78

Stanza	0.69	0.82	0.79	0.76	0.79	0.76	0.79
NB	0.61	0.87	0.83	0.77	0.78	0.81	0.82
BERT	0.75	0.90	0.89	0.85	0.87	0.86	0.88

Table 6.3 shows that BERT substantially outperformed the other sentiment classifiers for several reasons. First is (1) bidirectionality, where, unlike other conventional ML-based language models which process text in either left-to-right or right-to-left, BERT adopts a bidirectional approach for extracting knowledge from the processed text. Both directions are considered while learning each word's context during training, allowing it to better capture the contextual information and relationships between words in a sentence. Another reason is (2) being based on transformer architecture, which enables BERT to process long-range dependencies in text efficiently. Since transformers utilise self-attention mechanisms, it helps the classifier focus on representative parts of the inputs, assisting the model to capture semantic relationships between words efficiently. The (3) contextual word embedding is another reason BERT usually outperforms conventional ML classifiers. Other classifiers such as NB, which, although as in Table 6.3, outperformed other unsupervised approach (TextBlob, VADER and Stanza), it adopts traditional word embeddings like Word2Vec and GloVe, producing static representations for word regardless of the context. In contrast, the BERT embedding mechanism regards word contextuality, capturing polysemy (multiple meanings of a word), handling text ambiguity via capturing intricate patterns in text and producing more nuanced representations.

Considering the studies in Table 6.1, our developed model deals with several concerns regarding the effectiveness and generalisability identified within the current MOOC SA models. This includes (1) using one single source of data (previous literature on sentiment classification in MOOCs was based on single platforms only, Coursera and edX being the most popular platforms (Bulusu and Rao, 2021)) and hence less generalisable with relatively low number of instances compared to our obtained dataset; (2) lower model outputs, where the majority of the surveyed models in Section 6.3 are based on 2-polar classifier (positive or negative); (3) disregarding important sentiment indicators, such as emojis and emoticons, during text embedding; (4) adopting relatively outdated text embedding mechanism and (4) reporting average performance metrics only, preventing the evaluation of model performance at the level of class (sentiment). Thus, MOOCSent contributes to the knowledge, by addressing these limitations. Another purpose of building MOOCSent was to label FutureLearn learners' textual data (comments and replies) with their estimated sentiments, which were later employed as input features in the third experiment in Chapter 7, for discussion forums-based paid certification prediction.

With regard to the different version of BERT (Cased and uncased), learners are expected to use upper-case typing especially when expressing their sentiment. However, based on using both methods for training the model, BERT Base performed better and was hence chosen to build MOOCSent. This is in line with previous BERT-base sentiment or other text classification tasks, where uncased BERT has performed similarly to or even better than the Cased BERT (Jahan et al., 2021; Chiorrini et al., 2021; Peluso, 2022). Cased BERT was used, and it achieved negative:0.62, neutral:0.95, positive:0.86, BA: 0.81, WF1:0.87, F1:0.84, Acc:0.87. It was excluded because it (1) takes more time for training due to higher variations of text representation (more vocabulary size as the model retains the original casing of words in the training data) thus time-consuming, (2) performs better than uncased BERT in some specific case-sensitive tasks where the case of words carries essential information such as Named Entity Recognition (NER) or part-of-speech tagging, (3) consumes more computational resources and in same time performed similarly to or even worse in SA tasks according to our experiment and previous experiments (Jahan et al., 2021; Chiorrini et al., 2021; Peluso, 2022).

6.6. Epilogue

This study aims to propose a cross-platform MOOCs sentiment classifier using almost 1.2 million human-annotated learners’ comments obtained from 633 MOOCs delivered via Coursera, FutureLearn, Udemy and Stanford University. The initial experiment employed four commonly used architectures for predicting sentiment (TextBlob, VADER, Stanza and NB). Next, we used a context-aware classifier (BERT) for improving the sentiment classification, which outperformed the preceding classifiers, achieving BA (0.85).

As it is seen from the data used in this experiment, learner sentiments generally change towards the end of the course, where learners’ posts tend to be neutral during the course (as in Stanford data) and more positive by the end of the course (as in Coursera, Udemy, and FutureLearn data). This indicates the significance of learner sentiments as inputs when analysing text generated by learners. Therefore, in the next chapter, we used MOOCSent for annotating learner posts (which is the ultimate objective of building the current sentiment classifier) over the course weeks before using the entire discussion forums data for predicting certification.

Chapter 7 : Forum-based Prediction of Paid Certification in MOOCs

7.1. Prologue

In this chapter we investigate if MOOC discussion forum-based data can predict learners’ purchase decisions (paid certification) using various conventional and deep learning classifiers. For our temporal (weekly) prediction, we used data extracted from discussion forums (comments, replies, and likes) besides MOOCSent-tagged sentiment features and POS tagging to predict paid certification in MOOCs, yielding promising accuracies of 74% on average. The findings of this experiment are expected to help in the design of future courses and the prediction of profitability of future runs.

7.2. Introduction

Interaction in MOOCs discussion forums has been an influencing factor for predicting various behaviours such as course completion and detection of needed instructor intervention (Cagiltay, Cagiltay and Celik, 2020; Alrajhi *et al.*, 2022; Bonafini, 2017). Through social interaction and engaging with others, learners are believed to have more successful learning experience and greater commitment to the course content (Ferguson and Clow, 2015). The activities of learners in MOOC discussion forums span various practices, including interacting with peers, replying to other’s expressed thoughts, asking questions about the course materials, and liking other learners’ comments. This provides learners with a rich source of learning in a fully asynchronous way compared to traditional learning (Diver and Martinez, 2015).

While data extracted from MOOC discussion forums is commonly used to solve many MOOC challenges including dropout and identifying learners who require assistance, analysing learners' forum interactions to predict certification remains limited. Also, various types of learners’ collected data, for example, learner demographics, time spent on course content, and clickstream activities, have been used for predicting MOOC paid certification. However, according to our survey in Chapter 3, discussion forums, which are considered a rich source of learners’ interaction, have never been used for early predicting paid certification in MOOCs. Thus, this chapter proposes a forum-based predictor of learners’ financial decisions (course certificate purchase) via answering the following research question:

- *RQ4: Can raw and computed data extracted from MOOC discussion forums predict paid certification for courses?*

This study aims to predict the paid certification decisions of learners using MOOC discussion forums on whether the learners will purchase the course certificates. We use multidisciplinary course data from the relatively unexplored platform of FutureLearn to temporally predict purchase of certification. To the best of our knowledge, our temporal method in predicting MOOC learners’ financial decisions (purchasing a course certificate) using learners’ discussion forums has never been applied before.

Unlike previous studies on certification, our proposed model aims to predict the financial decisions of learners on whether to purchase the course certificate. Also, our work is applied to a less frequently studied platform, FutureLearn, as shown in Table 3.2. Another concern we address is study size, with 6 out of the total nine studies conducted on one course only. As learners may behave differently based on the course attended, previous models generalisability is unclear. Instead, we used a variety of courses from different disciplines, namely literature, psychology, computer science, and business.

Our SLR shows that literature has seldom used discussion forums extensively for the prediction of certification attainment. More specifically, textual data and generated features such as POSs and sentiments have not been used for predicting paid certification in relatively little studied platforms like FutureLearn.

7.3. Methodology

7.3.1. Data Collection

During their learning journey, learners can interact with each other in the form of expressing their thoughts (comments), replying to others' comments, and liking or disliking other learners' textual inputs. Unlike other activities that are typically mandatory towards the eligibility for course certificate attainment such as step access and question answering, learner participation in discussion forums is completely optional. This indicates more openness and learner autonomy in terms of the type of activities learners conduct on MOOC platforms.

The data used in this experiment were obtained from the discussion forums of the five FutureLearn topic-diverse courses (BD, BIM, SC, SP, and TMF). Table 7.1 shows the number of both non-paying and certificate-purchasing commenters compared to the total number of learners in each course. It can be seen that certificate purchasers were more interactive (almost 70% of certificate purchasers have at least posted a comment or replied to a comment), whereas non-paying commenters mark only around 11% of total number of non-paying learners.

Table 7.1. The number of non-paying learners (NLs), non-paying commenters (NL Commenters), certificate purchasers (CPs) and certificate-purchasing commenters (CP Commenters) in the 5 FutureLearn courses.

Cours e	#Comment s	#NLs	#NL Commenters	#CPs	#CP Commenters
BD	20,857	33,427	2,623	265	140
BIM	57,044	48,771	8,331	670	485
SC	6,013	5,808	484	69	37
SP	148,742	51,842	7,649	500	405
TMF	90,681	93,601	8,304	314	180
Total	323,337	233,449	27,391	1,818	1,247

7.3.2. Data Preprocessing

Preprocessing of the numerical features involves standard data manipulation such as replacing missing values to prepare the data for training. With regard to the textual features (comments and replies), regular expressions were used to remove unwanted characters, such as HTML/XML, punctuations, non-alphabet characters, etc., which are generally applied to filter out most unwanted text. Additionally, stopwords removal, lowering the cases of characters reforming contractions into the original words and grammar correction were also conducted. UNICODE_EMOJI and EMOTICONS_EMO lexicons were also used on this dataset to convert emojis and emoticons into tokenisable inputs, that is, converted into corresponding explanatory words.

7.3.3. Weighing Scheme

Term frequency-inverse document frequency (TF-IDF), which is the most common weighing scheme for text tokenisation, was used in this experiment. Unlike standard vectorisers, TF-IDF assigns a weight to each term in a document based on its frequency in the document (comment/reply) and a corpus of documents and generates a matrix of tokenised texts. After that, the vectors that represent the text can be used as input features for prediction. TF-IDF is calculated via two scores as follows:

$$\begin{gathered} tf(t,d) = log\ (1 + freq(t,d)) \\ idf(t,D) = log\left(\frac{N}{count(d \in D{:}\, t \in d)}\right) \\ tfidf(t,d,D) = tf(t,d).idf(t,D) \end{gathered} \tag{7.1}$$

Where, *t*: a term in a document (comment/reply) *d* and *D*: the whole corpus.

7.3.4. Part of Speech (POS) Tags

An example of further linguistic features such as POS tags is depicted in Figure 7.1, which too were measured based on the learners' comments/replies and included in the input features. This helps (with other features such as word and character counts) disclose any linguistic patterns associated with our binary-target classification (non-paying commenters and certificate purchasing commenters).

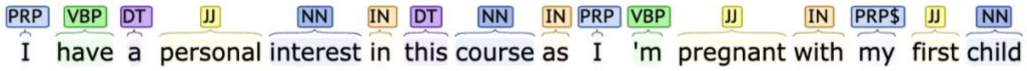

PRP VBP DT JJ NN IN DT NN IN PRP VBP JJ IN PRP$ JJ NN
I have a personal interest in this course as I 'm pregnant with my first child

Figure 7.1. An example of POS tagging for a BIM comment.

We used the Penn treebank POS tagger, which was built on a corpus comprising over 4.5 million English words (Marcus, Santorini and Marcinkiewicz, 1993). After merging both discussion forums-based raw and computed features, Table 7.2 was prepared, which shows the raw and computed features analysed in this study.

Table 7.2. The features (per week) utilised for predicting course certificate attainment.

Comment-based Features	Reply-based Features
#Comments	#Replies
% Positive comments posted	% Positive replies posted
% Neutral comments posted	% Neutral replies posted
% Negative comments posted	% Negative replies posted
#Likes received	#Likes received
Word count	Word count
Character count	Character count
tf-idf	tf-idf
POS tags	POS tags
	% Positive replies received
	% Neutral replies received
	% Negative replies received

7.3.5. Data Resampling

Considering the highly imbalanced dataset we used in this study, where the output classes ratio on average is around 22:1 as stated in Table 7.1, we oversampled the training data only using synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002). SMOTE is a commonly used minor class oversampling method via generating synthetic samples of the minor class. Overcoming the overfitting issue, which typically occurs when random oversampling technique is adopted, SMOTE emphasises the minor class space and interpolates new instances similar to the original instances in terms of feature space. The validation architecture followed in this experiment is hold-out or train and test (T/T) split with a ratio of 70:30 for training and test data. As SMOTE generates synthetic instances of the minor class, using CV may result in the same instance being simultaneously in the training and test set; thus, T/T is the proper validation architecture to use in the present experiment. With the statistics in Table 7.1 in mind, the overall test data size for assessing the model performance was 8591 instances.

7.3.6. Model Architecture

The current experiment applied 4 deep classifiers: CNN, RNN, LSTM, and GRU, explained earlier in Section 4.4.4. Given that we employed both textual and numerical discussion forum-extracted features, we adopted these models for their ability – compared to other conventional ML models - to handle multi-inputs.

The first layer of the model is the embedding layer – an embedding matrix – which performs as a lookup table of textual inputs (comments and replies) and converts the text_input into a dense representation. This layer maps each learner's textual input representation (X) into a matrix of size: $l_c \times d : X \in R^{l_c \times d}$, where *l* the the maximim length of textual inputs in course *c*, which equals 90 percentile of the word count.

In the CNN architecture, the output of the embedding layer is fed into a 1D CNN layer to process the text_input using *ReLU* as the *activation function*. Next, a *flatten layer* is applied to convert arrays into 1D vector. Next, a dropout layer is applied – with 0.5 value as commonly set (Srivastava et al., 2014) – to reduce overfitting via setting input units to 0 with a frequency of rate at each step during training. The *dropout layer* is connected to a *dense layer* which is connected to a *concatenate layer*. The numerical feature inputs are also connected to the *concatenate layer*, which is connected to another *dense layer*. The final vector obtained is then fed into the *output layer* for classification. The other models of RNN, LSTM, and GRU followed a similar structure, apart from the second layer which was structured based on the corresponding model. Figure 7.2 depicts the overall workflow of our predictive model.

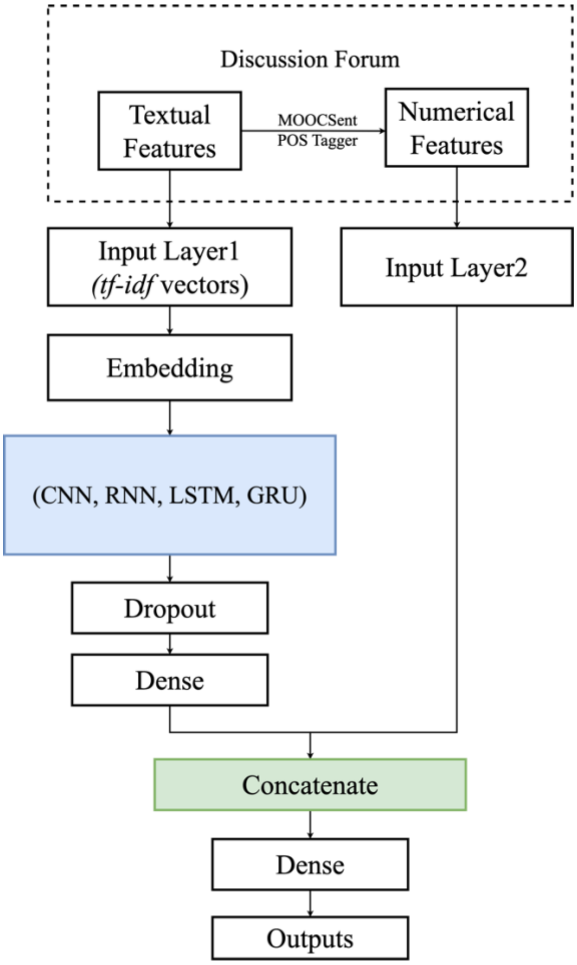


Figure. 7.2. General workflow illustration of the predictive model.

7.3.7. Automated Hyperparameters Optimisation

To minimise the time spent on manually optimising the hyperparameters of the predictive models, automated search for the optimal values of the hyperparameters was adopted. We used the Optuna (Akiba et al., 2019), a define-by-run framework, which dynamically searches and prunes the optimal parameters. Unlike previous hyperparameters tuning frameworks such as SMAC (Hutter, Hoos and Leyton-Brown, 2011), Vizier (Golovin et al., 2017), Autotune (Koch et al., 2018), and Hyperopt (Bergstra et al., 2015), Optuna enables a user to dynamically construct the search space more dynamically compared to previous frameworks. Thus, the effectiveness of optimisation is improved with Optuna by combining the efficient searching and pruning algorithm.

7.3.8. Dealing with Bias

Data-based algorithmic bias in educational predictive models has been observed in several studies, which consequently emphasised the need for more representative data as one of the approaches to mitigate potential sources of unfairness (Baker and Hawn, 2021). While the dataset adopted in this experiment was obtained from one platform and only contains learners' interactions in the course discussion forums, it shows considerable representation. The dataset spans various runs (23) of 5 different MOOCs, covering 4 distinct disciplines (literature, psychology, computer science, and business), allowing to longitudinally predict paid certification based on a rich source of data. As shown in Table 3.2, most of the surveyed works were based either on one course or run/iteration. Another novelty of the present experiment is predicting paid certification in MOOCs at an early stage (starting with using data from the first week of the course only), which would be impossible without the high representation of the datasets used in this experiment.

Considering the highly imbalanced dataset we used in this study, where the output classes ratio on average is around 22:1, as stated in Table 7.1, we oversampled the training data only using the Synthetic Minority Oversampling technique (SMOTE) (Chawla et al., 2002). SMOTE is a commonly used minor class oversampling method via generating synthetic samples of the minor class. Overcoming the overfitting issue, which is highly associated with bias and typically occurs when a random oversampling technique is adopted, SMOTE emphasises the minor class space and interpolates new instances similar to the original instances in the feature space. The "need for more data for a higher level of model generalisation and further validating the achieved results" was the most stressed call by the surveyed studies. As demonstrated earlier in Figure 3.7, a considerable number of studies (n = 14/25) have based their findings on a few courses (from one to three courses only); thus, it is challenging to consider the findings of these models generalisable. Furthermore, learners' behaviours and certification rates differ based on the subject and discipline of the MOOC (Cobos and Jurado, 2018). Therefore, building the model on a diverse dataset would help increase the findings' generalisability and the reliability of the results.

One common procedure for mitigating algorithmic bias is comprehensively finetuning the algorithms' metrics, which can help the model perform reasonably on all subcategories of the outputs (Baker and Hawn, 2021). Although this procedure may reduce the model's overall accuracy (Xiang et al., 2022), it allows for more harmonious performance, i.e., BAs, indicating a fair consideration of all the output classes. Although configuring predictive algorithm parameters is essential for improving model predictability and fairness, most surveyed works in Chapter 3 (please see section 3.6.2.2 for more details) skipped this step, or simply let the model assign the default parameters. Having the parameters tuned generally helps the model achieve better forecast results and find and diagnose common modelling issues such as bias, underfitting, and over-fitting.

To address this issue, an automated search for the optimal values of the hyperparameters was adopted to (1) reduce any potential chance for bias and (2) minimise the time spent on manually optimising the hyperparameters of the predictive models. We used the Optuna (Akiba et al., 2019), a define-by-run framework which dynamically searches and prunes the optimal parameters. Unlike previous hyperparameters tuning frameworks such as SMAC (Hutter, Hoos and Leyton-Brown, 2011), Vizier (Golovin et al., 2017), Autotune (Koch et al., 2018), and Hyperopt (Bergstra et al., 2015), Optuna enables a user to dynamically construct the search space more dynamically compared to previous frameworks. Thus, the optimisation effectiveness is improved with Optuna by combining the efficient searching and pruning algorithm.

It is worth noting that while the present experiment analysed in-depth learner's interactions and textual inputs in the discussion forums to predict paid certification, *which has not been done before*, it may impose some bias. Table 7.1 shows that the number of NL Commenters was 27,391, which marks only around 12% of non-paying learners. However, the statistics are better for CP Commenter (1,247), where around 69% of certificate purchasers had at least 1 comment/reply posted, thus included in the analysis.

7.4. Results

The present experiment is aimed at the fourth research question, of whether raw and computed data extracted from MOOC discussion forums can predict paid certification. The raw data (learner textual inputs) from learner discussion forums, sentiment classification using MOOCSent, and computed features (number of likes received for each textual input), in addition to several features extracted from the texts (e.g. character counts, word counts, and part of speech (POS) tags for each textual instance) were used to build the predictive model.

This section shows the performance of our temporal predictive model using the first-week data only (Table 7.3) and the first half of the course (Table 7.4) where Rec_0 = recall score for non-paying learners (NLs or Class 0) and Rec_1 = recall score for certificate purchasers (CPs or Class 1). Additionally, a comparison of the classifiers' performance using BA, which is the most representative performance metric in our experiment, using the first-week data only (BA1) and the first half of the course (BA2) distributed by courses and classification algorithms is provided in Figure 7.2. In Tables 7.3 and 7.4, BA, WF1 and Acc were also reported to report the overall performance of the model. BA is defined as the average of recall obtained on each class, which equals the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate), whereas WF1 can be considered as a weighted trade-off between precision and recall. For Acc, it calculates the model's overall performance, giving the same weighting for all classes by dividing the total number of correctly predicted instances by the total number of instances.

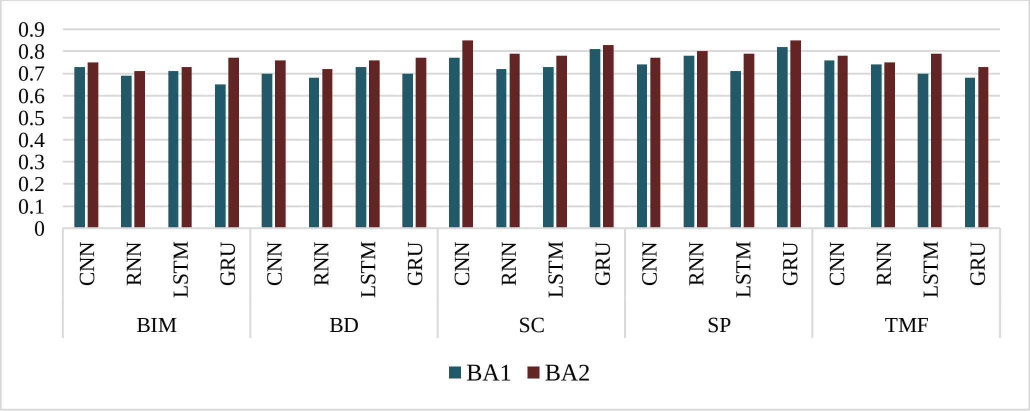


Figure. 7.3. Learner classification results (BAs) using the first-week data only (BA1) and the first half of the course (BA2) distributed by courses and classification algorithms.

The overall performance of the four employed algorithms achieved good Recall and Balanced accuracy with the first-week data only (average BA = 0.72), with considerable improvement when further features of the first half of the course were included (average BA = 0.77). Regarding the best classifiers, all four algorithms achieved similar performance, with CNN and GRU being equally the best in the first-week-only scenario and GRU being the leading classifier over the mid-course scenario. One considerable improvement at the class level was Rec_1, which reached 0.80 over the mid-course from 0.69 over the data for the first week only.

The results in Tables 7.3 and 7.4 show that Rec_0 has decreased in 12 instances across the five courses. In these specific cases, the activities of both groups (NLs and CPs) were similar in the first week, which directed the classifiers towards one class (Class 0/NLs), resulting in lower Recall for the other class (CPs). However, while the decrease was by only around 1% on average (from 0.76 to 0.75), Reac_1 has jumped by 11% from 0.69 to 0.80, indicating a better performance of the predictive model

(BAs). This has resulted in higher BAs, which, on average, also improved (from 0.73 in the first week-only experiment to 0.77 in the first half of the course). Therefore, the overall Recall of all classes should always be considered when assessing a model's predictability.

Table 7.3. Learner classification results distributed by course using the first week data only; Rec_0 = recall score for non-paying learners (Class 0) and Rec_1 = recall score for certificate purchasers (Class 1).

Course	Classifier	Rec_0	Rec_1	BA	WF1	Acc
BIM	CNN	0.74	0.71	0.73	0.81	0.74
	RNN	0.70	0.69	0.69	0.78	0.70
	LSTM	0.75	0.67	0.71	0.82	0.74
	GRU	0.63	0.67	0.65	0.73	0.63
BD	CNN	0.79	0.60	0.70	0.84	0.78
	RNN	0.75	0.61	0.68	0.81	0.74
	LSTM	0.86	0.60	0.73	0.88	0.84
	GRU	0.73	0.67	0.70	0.81	0.73
SC	CNN	0.78	0.75	0.77	0.85	0.72
	RNN	0.79	0.64	0.72	0.83	0.78
	LSTM	0.86	0.60	0.73	0.88	0.85
	GRU	0.76	0.86	0.81	0.83	0.77
SP	CNN	0.84	0.64	0.74	0.88	0.83
	RNN	0.74	0.82	0.78	0.82	0.74
	LSTM	0.82	0.60	0.71	0.85	0.81
	GRU	0.68	0.96	0.82	0.78	0.69
TMF	CNN	0.70	0.82	0.76	0.80	0.70
	RNN	0.75	0.72	0.74	0.84	0.75
	LSTM	0.75	0.65	0.70	0.83	0.74
	GRU	0.75	0.60	0.68	0.84	0.75

Table 7.4. Learner classification results distributed by course using the first half of the course; Rec_0 = recall score for non-paying learners (Class 0), Rec_1 = recall score for certificate purchasers (Class 1).

Course	Classifier	Rec_0	Rec_1	BA	WF1	Acc
BIM	CNN	0.82	0.68	0.75	0.86	0.81

	RNN	0.67	0.75	0.71	0.77	0.67
	LSTM	0.67	0.78	0.73	0.77	0.68
	GRU	0.79	0.76	0.77	0.84	0.79
BD	CNN	0.69	0.83	0.76	0.78	0.70
	RNN	0.75	0.69	0.72	0.82	0.75
	LSTM	0.69	0.83	0.76	0.78	0.70
	GRU	0.67	0.86	0.77	0.77	0.68
SC	CNN	0.90	0.80	0.85	0.92	0.90
	RNN	0.77	0.80	0.79	0.84	0.77
	LSTM	0.85	0.71	0.78	0.87	0.84
	GRU	0.86	0.80	0.83	0.89	0.86
SP	CNN	0.74	0.80	0.77	0.81	0.74
	RNN	0.73	0.87	0.80	0.81	0.74
	LSTM	0.70	0.89	0.79	0.79	0.71
	GRU	0.72	0.97	0.85	0.81	0.73
TMF	CNN	0.71	0.84	0.78	0.81	0.72
	RNN	0.72	0.78	0.75	0.82	0.73
	LSTM	0.69	0.88	0.79	0.79	0.69
	GRU	0.77	0.69	0.73	0.85	0.76

One of the reasons for clickstreams to be better predictors for paid certification is that they contain mandatory activities that learners, especially certificate seekers, need to perform to be eligible for certificate attainment. For instance, the average score in edX and FutureLearn that learners need to achieve for certificate eligibility is 65%. These requirements incentivise certificate purchasers to interact more with the course content, thus rendering more distinct data compared to non-paying learners. However, in discussion forums, participation is completely optional and – unlike the successive completion markings of the learning steps – can be performed at any time of the course. This resulted in a noisier and smaller dataset, as explained in Section 7.3.1 and consequently affected the predictive model performance.

Both word-level and sentence-level (weekly inputs) were analysed in this thesis, based on the scenario of the experiment. In this specific experiment, the textual inputs are the sum of various inputs of a learner (comments/replies) with a certain scenario (first week only or the first half of the course).

Consequently, further computed features at the sentence level, besides weighing the text with *tf-idf*, were performed to enrich the inputs. This included standard computed features such as characters and word counting as well as POS tagging, which marks each word in the learner's textual inputs based on its corresponding POSs at the sentence level.

Using the Optuna framework, various DL parameters were automatically optimised in this experiment, including the model optimiser, learning rate, and the number of epochs. Optuna dynamically – based on the *define-by-run* principle – constructs the search space for the best set of parameters, maximising the number of trials conducted with less manual intervention. Thus, the effectiveness of optimisation is improved with Optuna by combining the efficient searching and pruning algorithm. With our fine-tuned parameters mentioned in section 7.3.7, the framework parallelises hyperparameter searches over multiple trials until the optimal result is reached.

This MOOC prediction task is considered highly challenging, compared to other MOOC tasks, such as predicting dropout, completion, and learner characteristics. The reason for this is the severe data imbalance of the binary class, where course certificate purchasers account for less than 1% of the total number of enrolled learners.

7.5. Epilogue

This study compared four deep classifiers to predict course purchasability using discussion forum data from five MOOCs. Our proposed model achieved various BAs, averaging 0.77, using only the first half of course data. Thus, it can predict relatively early whether or not a learner will purchase a certificate at the end of the course based on their discussion forum-based interaction.

Chapter 8 : Discussion

8.1. Prologue

This chapter discusses the SLR outcomes and opportunities for future development in the field of certification prediction in MOOCs. Additionally, it explores the statistical and ML-based results of this thesis and the novelty and limitations based on the experiments conducted in Chapters 5, 6 and 7. It also elucidates how these results contribute to the knowledge considering the research gaps identified within the surveyed literature in Chapter 3.

8.2. Introduction

As stated earlier in Chapter 2, MOOCs were developed specifically to reach an unlimited number of potential learners, resulting in attracting the attention of millions of learners across the entire educational landscape. However, later, with these platforms becoming more independent educational companies, monetised content became necessary to fund MOOCs and ensure the sustainability of such platforms. This trend has resulted in the monetisation of standard courses and included new forms of monetised content, such as micro-credentials, corporate training, and university degrees.

Despite the unparalleled success of MOOCs, the staggeringly decreasing certification rates and, more critically, the decline in the paid certification statistics of a given course over various runs/iterations, are considered great threats to platform sustainability.

The present thesis attempts to address this issue by firstly surveying the literature on predicting certification in MOOCs and identifying the limitations within the surveyed studies.

Secondly, by adopting sequential steps through using learners' clickstreams, reviews, and comments from various MOOC platforms and learners' discussion forum interactions, it attempts to address the issue of predicting paid certification, and thus potentially helping both MOOC providers, as well as, indirectly, learners, based on various data available from a MOOC course, as well as at various time points of a course.

Additionally, the contributions of the present thesis include building the sentiment classifier of MOOCSent, to predict sentiment in a massive dataset of around 1.2 million learners' comments. The purpose of building MOOCSent was to use the colllected labelled 1.2 million MOOC reviews to train the model and then label FutureLearn textual data (comments and replies) with their estimated sentiments.

These classified sentiments were later employed as input features in the third experiment, for discussion forums-based certification prediction, since the text dataset used in this specific experiment is unlabelled with learners' sentiments, which have been noted as ideal determinants of learner success in MOOCs (Sraidi *et al.*, 2022; Wen, Yang and Rose, 2014; Chaplot, Rhim and Kim, 2015; Dalipi, Zdravkova and Ahlgren, 2021).

8.3. Systematic Literature Review

Our SLR contributes to presenting a promising synthesis of the state of the art on this research topic, considering the struggle of MOOC platforms to build their own business models in addition to the recent transition, since 2017, towards paywalled content like micro-credentials, corporate training, and online degrees with affiliate university partners. It also serves as a roadmap for the multidisciplinary community of researchers in the educational domain (e.g. data scientists, statisticians, and educators) to explore the prediction of certification in MOOCs from a wider angle.

The survey in Section 3 identified the diverse methodologies followed and evaluated their applicability in a real-world scenario. Also, this survey categorised the surveyed models based on several aspects: their utilised input features, size, types of data used and their sources, model generalisability, prediction methods, and performance metrics reported. Additionally, several critical methodological concerns, as further discussed in Section 3.7, were highlighted, including (1) the need for more data for a higher level of model generalisation and further validation of the achieved result; (2) sample extensive filtration, which was found positively correlated with including more features and consequently affecting model generalisability; (3) the insufficient experiment elaboration on essential parts of experiments, for example, feature engineering and selection; (4) parameters fine-tuning, which was little reported within the surveyed works, (5) non-realistic modelling, where some models followed methodologies that were not actionable in real-life scenarios; and (6) the scarcity of models that could predict paid certification early. The present thesis addressed these limitations and methodological concerns from various aspects as discussed within the following subsections.

8.4. Clickstream-based Prediction

The results in section 5.4 firstly answered the first research question in Section 1.3 by exploring how our processed features (access, attempts, correct answers, and wrong answers) can temporally distinguish course certificate purchasers from non-paying learners based on their activity data. Our temporal analysis showed high statistical significance at various levels when comparing the behaviours of non-paying learners with those of certificate purchasers across the 5 courses analysed. Tables 5.5-5.8 in Section 5.4.1 show the statistical analysis results using the Mann–Whitney U test (of the 3 time points: first week, mid-week, and last week) in each course. As the courses analysed spanned over different numbers of weeks, we have selected the first, middle, and last weeks to report the results. Given below are the results for the four analysed activities (access, attempts, correct answers, and wrong answers).

For courses with an even number of weeks, we have selected the middle week closer to the start of the course for analysis. Our analysis indicated that paying learners were generally more engaged with the course content, in terms of accessing the content more frequently, attempting more questions, and answering more questions correctly, and reattempting more questions answered incorrectly. Considering platforms allow learner several attempts to answer a question correctly, the later (number of wrong answers) indicates certificate purchasers' persistence to reach the minimum score required to be eligible for the course certificate. While all the results of the statistical analysis were very significant ($p < 0.001$, ranging from 4e-23 to 0), the test showed more significance towards the end of the course. Course-wise, the difference in the activities of both group of learners in SH was the most significant, whereas the significance of these 4 predictors based on the results of the last week can be placed in a descending order as attempts, correct answers, wrong answers, and access. Thus, non-paying course takers behave differently from course purchasers as to their activities of access and answering questions (attempts, correct answers, and wrong answers).

The second part of the results in Chapter 5 answers our second research question on whether the learner clickstreams can be used to predict paid certification in MOOCs. The results achieved a promising BA, ranging from 0.77 to 0.95, across the 5 domain-varying courses. The classifiers performed differently based on the course analysed, where SVC performed the best in BIM, ET in BD and SP, and LR in CS and TMF. In general, the improvement in the performance of the classifiers was lower towards the end of the courses compared to the difference between the first week only and the first half of the course. This may indicate that course purchasers exert more effort until they reach the minimum requirements for certification (typically just after the middle of the course). Thereafter, the level of interest in the course content in terms of access, question answering, and time spent learning is reduced; hence, at this stage, activities that are more similar to non-paying learners are performed even by the paying course takers.

As discussed in Chapter 3, the correlation between the time spent on the course content and certification was statistically analysed by various previous works including those by Cobos and Jurado (2018) who used the learners’ time spent on assignments and videos and Goli, Chintagunta and Sriram (2019) who used the total time spent (minutes) and the average session duration (minutes). Tian *et al.* (2017) analysed the correlation between the time spent by learners on content in days and certification, whereas Qiu *et al.* (2016) examined the impact of effective learning time spent and certification attainment. However, the time spent on content as a feature for building a paid certification predictive model has not been used before on MOOC data in a large scale. This highlights the contribution of the present experiment which computed this feature of a relatively large data of 23 runs of 5 courses.

8.5. MOOCSent Sentiment Classifier

This experiment answers the third research question stated in Section 1.3 by building a cross-platform MOOCs sentiment classifier using over 1.2 million human-annotated learners’ comments and reviews obtained from 633 MOOCs. We used various lexicon-, ML-, and DL-based classifiers to evaluate the performance of our model. The ultimate objective of building this classifier was to annotate our learner’s unlabelled comments in the dataset of FutureLearn posted weekly by the learners and use these with other input features for predicting paid certification based on learners’ interactions in discussion forums.

Table 6.3 shows the results of our sentiment prediction model using TextBlob, VADER, Stanza, NB, and BERT, and it can be seen that the classifiers achieved different BA (average 0.76) ranging from 0.72 with TextBlob to 0.85 with BERT. Unsurprisingly, BERT outperformed lexicon and conventional ML-based classifiers due to its context-aware learning. This could be attributed to the fact that BERT uses text contextuality to capture representative knowledge from the processed text and thus detects the true learner’s sentiment and performs better compared to the other models.

At the level of class (sentiment), the models were able to equally detect the neutral and positive classes better than the negative class. This is perhaps due to the nature (size) of the training data; nevertheless, it was augmented by replacing words with their synonyms of the same POS from the WordNet thesaurus as can be seen in Table 6.2. Model-wise, TextBlob seems more sentiment-sensitive given that it is the only model that was able to detect both sentiment polarity (negative and positive) more than the neutral class.

Regarding parameters fine-tuning, we observed a high correlation between the computational cost and the model being fine-tuned. For instance, the BERT-based experiment in MOOCSent in Chapter 6 requires massive resources to use BERT where each run took around 17 hours to be completed. Thus, parameters were fine-tuned based on literature recommendations to reduce the computational cost and at the same time reach an optimal performance. However, in the last experiment, which was also based on DL algorithms but required relatively lower resources due to the size of the data used, an automatic fine-tuning (Optuna) was adopted.

Regarding the selection of classifiers in Chapter 6 and 7, this step was based on the nature of the data used in each experiment. In MOOCSent, we adopted the most common lexicon-based and ML models to build the model. Next, we adopted BERT to further evaluate the level to which contextual architectures may improve the model performance. In Chapter 7, we adopted the most common DL architectures as our data contained two types of features, namely numerical and textual. The adopted models (CNN, RNN, LAST, and GRU) are commonly used for handling datasets of two or more inputs.

The validation architectures adopted in our experiments were based on the nature of data analysed and the preprocessing conducted. In the first experiment, we used the stratified cross-validation (CV) technique, which uses *k* folds (portions) of the data, preserving the same percentage of samples for each class in each fold to train and test the model on different iterations. The cross-validation architecture is generally better, as it avoids overfitting, by allowing the model to train on multiple train-test splits. Consequently, a better estimation of the model performance on unseen data is indicated. The other validation architecture is hold-out or train and test (T/T) split. This strategy was followed in MOOCSent experiment, as the Stanford dataset consists of over-the-course comments. Thus, we used it for testing the model performance, because the ultimate purpose of building the model was to annotate the FutureLearn comments and replies with sentiment for the final experiment. The T/T split was also adopted in the last experiment due to oversampling the training data. Considering this experiment contains a relatively lower number of instances of both classes, oversampling of the training data (using SMOTE) helped the model to improve its performance. As SMOTE generates synthetic instances of the minor class, using CV may result in the same instance to be in the training and test set at the same time; thus, T/T is the proper validation architecture to use in this case.

Unlike most of the reported studies, our predictive model MOOCSent targets three (not only two) polarities: positive, negative and neutral, being able to capture more sentiment classes of learners (please refer to the performance of the model on the three classes in Table 6.3). A further extension could look into more detailed categories of sentiment. This was not further explored in this thesis, as the combined predictor outperformed the competition at the time, but could be considered for futher optimisation work.

8.6. Discussion Forum-based Prediction

This experiment is aimed at the fourth research question on whether raw and computed data extracted from MOOC discussion forums can predict course paid certification. In our final experiment (in Chapter 7), the raw data (learner textual inputs) from learner discussion forums, sentiment classification using MOOCSent, and computed features (number of likes received for each textual input), in addition to several features extracted from the texts (e.g. character counts, word counts, and part of speech (POS) tags for each textual instance) were used to build the predictive model.

The overall performance of the four employed algorithms has achieved good recall and balance accuracy with the first week data only (average BA = 0.72), with considerable improvement when further features of the first half of the course were included (average BA = 0.77). Regarding the best classifiers, all the four algorithms achieved similar performance, with CNN and GRU being equally the best in first week-only scenario, and GRU being the leading classifier over the mid-course scenario. One considerable improvement at the class level was Rec_1 which reached 0.80 over the mid-course from 0.69 over the data for the first week only.

One of the reasons for the clickstreams to be better predictors for paid certification is that they contain mandatory activities that learners, especially certificate seekers, need to perform to be eligible for certificate attainment. For instance, the average score in edX and FutureLearn that learners need to achieve for certificate eligibility is 65%. These requirements incentivise certificate purchasers to interact more with the course content, thus rendering more distinct data compared to non-paying learners. However, in discussion forums, participation is completely optional and – unlike the successive completion markings of the learning steps – can be performed at any time of the course. This resulted in a noisier and smaller dataset as explained in Section 7.3.1 and consequently affected the predictive model performance.

Both word-level and sentence-level (weekly inputs) analysis were adopted in this thesis based on the scenario of the experiment. For instance, MOOCSent uses textual inputs only to generate outputs (sentiment). Thus, word-level embedding (using the lexicon-based analyser and BERT tokeniser) was adopted. This is in line with adopting the most promising classifiers for that specific task, that is, sentiment classification. However, in the last experiment (in Chapter 7), the textual inputs are the sum of various inputs of a learner (comments/replies) with a certain scenario (first week only or the first half of the course). Consequently, further computed features at the sentence level, besides weighing the text with *tf-idf*, were performed to enrich the inputs. This included standard computed features such as characters and word counting as well as POS tagging, which marks each word in the learner's textual inputs based on its corresponding POSs at the sentence level.

Using the Optuna framework, various DL parameters were automatically optimised in this experiment including the model optimiser, learning rate, and the number of epochs. Optuna dynamically – based on *define-by-run* principle – constructs the search space for the best set of parameters maximising the number of trials conducted with less manual intervention. Thus, the effectiveness of optimisation is improved with Optuna by combining the efficient searching and pruning algorithm. With our fine-tuned parameters mentioned above, the framework parallelises hyperparameter searches over multiple trials until the optimal result is reached.

8.7. Potential Algorithmic Bias

Algorithmic bias within the context of educational predictive models has been identified, shedding light on various facets of this increasing issue. Initially, it was noted that algorithms are often perceived as objective and fair, but a growing body of evidence suggests that they can inadvertently incorporate biases, leading to unfair outcomes for certain groups of learners. A recent survey shows that demographics specifically emerged as a critical factor associated with algorithmic bias in education, encompassing gender identity, race, nationality, ethnicity, age, national origin, and sexual orientation (Baker and Hawn, 2021). However, the research dataset employed in this thesis excluded demographic information, to develop a more generalisable predictive model for paid certification, thus avoiding extensive sample filtration that has plagued previous predictive models, as discussed in section 3.7.2.

Our experiments employed several strategies to mitigate potential bias, including representative data preprocessing techniques and algorithmic metrics. These strategies included data shuffling, to expose the model to diverse learner activities, stratified cross-validation, to prevent overfitting, and oversampling of minority classes using the Synthetic Minority Oversampling Technique (SMOTE) to address the imbalanced dataset. Additionally, we used data labelled by learners aimed to enhance annotation accuracy and fairness in results. Key sentiment indicators, such as emojis and emoticons, were also appropriately encoded, to ensure more reliable and fairer results. The dataset was highly representative, comprising data from diverse MOOC platforms, disciplines, and runs. This diversity allowed for the longitudinal prediction of paid certification based on rich data sources, a notable contribution to the field. Moreover, early-stage prediction of paid certification was made possible due to the dataset's high representation power.

An automated search for optimal hyperparameter values using the Optuna framework was employed, to finetune algorithmic parameters automatically. This approach aimed to reduce potential bias and minimise manual optimisation efforts, while improving model predictability and fairness. In conclusion, this study comprehensively addressed algorithmic bias in educational predictive models, offering insights into data collection, preprocessing, and algorithmic tuning strategies to promote fairness and reliability. The research contributes to the ongoing efforts to mitigate bias in MOOC predictive modelling, by emphasising the importance of representative datasets and thoughtful algorithmic design.

8.8. Limitations

While the present thesis contributes novel knowledge in the field of predicting paid certification in MOOCs, it contains some limitations as is true for any research. Highlighting these limitations is expected to have a great impact on defining the directions for future improvement.

- While MOOC providers have been going through various tiers of content monetisation (Figure 2.4), it is inevitable to engage learner data via research collaboration to achieve this goal in a timely manner. Our endeavours to obtain access to more data for further validation and generalisation of the findings of our experiments included formal requests submitted to many platforms, including Noon Academy (March 2019); Rawaq, KKUx, and Edraak (October 2019); and Coursera, Udemy, and edX (November 2019). All of these research collaboration proposals were rejected, except for Rawaq and Edraak who shared the data of three courses (51,804 learners) and two courses (2,377 learners), respectively, after several meetings. However, the data shared did not contain any information on certification. This again indicates the sensitivity of learners' financial records and reservations of the platforms on sharing such data outside the organisation boundaries.
- One of the limitations of our NLP-based analysis in Chapter 6 and 7 is that it dealt with English-only texts. This has been restricted by the language-specific NLP tools used for data preprocessing (such as text augmentation by paraphrases replacement in Chapter 6 and POS tagging in Chapter 7) which currently deal with English-only data. However, platforms based in English-speaking countries only form one fifth of the total number of providers worldwide (see Appendix A). This encourages us to explore other international platforms in the future and analyse how learners' behaviour, specifically in terms of certification attainment, may differ.
- It is worth mentioning that the finding of this thesis, although they were based on courses of different disciplines, may be platform specific. One instance is the statistics on the course loss enrollees and, consequently, purchasers, over the consecutive runs of the courses analysed. Although this decline in certification rate is a common issue with MOOCs due to the transition towards paid content (Cagiltay, Cagiltay and Celik, 2020), findings of our research are still subject to careful interpretation.
- Another point to consider is that while different types of data (clickstreams and the features extracted from discussion forums) have been employed to predict paid certification for

courses, other types including pre- and post-course surveys and demographics can also be used (as stated in our SLR in Chapter 3) and they may yield promising results. However, regarding these two data types specifically, they are either rarely collected from many learners as responses are typically not mandatory or subject to response biases (Kizilcec and Halawa, 2015). A study of a massive dataset of 70 MOOCs found that only 25% of learners completed the demographic survey, and only around 2% earned certificates at the end of the course, hence rendering the portion of certified learners with known demographics very small (Goli, Chintagunta and Sriram, 2019). While platforms compete to lower entry barriers and ease restrictions on access to courses, they should consider making demographic survey completion mandatory for at least course purchasers so as to obtain more meaningful data from learners.

8.9. Future Works

The limitations discussed in Section 8.7 directly propose areas for development. Given below is a list of future works planned to improve the current thesis outcomes.

- The last two years, coinciding with the spread of COVID-19 and the resulting lockdown of educational institutions and travel restrictions in many countries, were exceptional in MOOCs history. In 2020, around one-third of the total number of learners since the emergence of MOOCs (60 million of 180 million learners) had joined (Shah, 2020), and, by the end of the year, the number of offerings had reached 16,300 courses, out of which 2,800 courses were launched within the year. This was in conjunction with the increase in the number of micro-credential courses offered that reached 360 from just 170 in 2019 and 19 new online degrees that were introduced (Shah, 2020). The following year also showed a further pandemic-fuelled increase in the popularity of MOOCs. As of 2021, 40 new million learners enrolled in at least one MOOC and over 3,000 new courses were launched (Shah, 2021a). This unprecedented transformation towards online learning, particularly in the case of MOOCs, appeals for a closer investigation of how learners progress in these courses and how MOOC learner success, in the form of obtaining a course completion certificate, can be modelled and predicted. Furthermore, there is a need for a comparative study to examine the extent to which learner activities, especially certification, has changed as a result of COVID-

19. We wish that access to the recent runs of the courses analysed in this thesis would be granted by the University of Warwick.

- Numerous sophisticated NLP models have been built based on the Transformers mechanism and transferer learning that have been enriching the NLP research society (Torrey and Shavlik, 2010). Besides BERT, Xlnet has been introduced subsequently which was trained on more data (over 130 GB of 32 billion words compared to 16 GB of 3.3 billion words in BERT) and achieved 2–20% improvement over BERT on different benchmark data (Yang *et al.*, 2019). We plan to further expand our experiment in the future using Xlnet because our current computational resources are not adequate to use Xlnet, which uses 10 times more data than BERT and thus requires more time and computational resources for execution.
- The sentiment annotator (MOOCSent) and POS taggers, which were built based on Penn Treebank of a corpus consisting of over 4.5 million English words, have been adopted to annotated English-only textual inputs in the discussion forums. However, considering MOOCs are learning platforms accessible worldwide, we noticed that some learners interact with each other in their own language. One of our next-step improvements is to expand our analysis to include further languages. This is considered especially for courses that are offered on FutureLearn by HEIs/instructors from non-English-speaking countries, such as the City University of Hong Kong and University of Malaya in Malaysia.

8.10. Future Research Direction and Opportunities

Based on the outcomes of our SLR and the methodological gaps highlighted in Section 3.7, future research direction and opportunities for improvement are discussed in this section. They are expected to help improve the transparency, performance, and scalability of the current MOOC certification predictive models.

- The meagre statistics on the number of enrolled learners, which range between only 46% and 60% of the registered learners, is an interesting phenomenon to be further analysed. MOOCs are generally well marketed, especially on social media, which yields a massive array of registered learners; nevertheless, the number of enrolled learners has shrunk massively compared to the number of registered learners. It is still valid to investigate whether the course introductory materials, where a potential learner lands, might be reengineered to convince more learners to commence their courses. Some motivational strategies to encourage learners to enrol, such as regular reminders, have been recommended and have already been employed by platforms (Cagiltay, Cagiltay and Celik, 2020). However, the instructional design of the introductory content of the courses might be part of the reason that need further investigation. In relation to this, Mullaney and Reich (2015) suggested that reminding learners of the general course objectives at the beginning of each week might help increase the retention rate throughout the course.

- Additionally, studying the impact of course prices on affordability seems to be an unexplored topic. Current studies have neither reported the impact of the course price on the certification nor explored the long-term impact of such an investment on learners' future endeavours. Despite the difficulty of obtaining such data (e.g. the unavailability of course prices when they were offered and the more challenging prices of discontinued courses or runs), we hope that researchers will be able to explore the presence of such long-term financial impacts.
- The non-replicability of the current certification predictive models is another concern to highlight, a common phenomenon in the educational technology field at large (Makel and Plucker, 2014). Replicability, especially with the rapid proliferation of different methods of modelling learner behaviours in MOOCs (Gardner and Brooks, 2018b), would help validate previously achieved results and find a helpful basis for comparing predictive certification models (Jiang, Fitzhugh and Warschauer, 2014). However, out of the 25 models surveyed, none either used a publicly accessible dataset or offered access to the dataset for future improvement to the proposed models by the broad community of educational data scientists. Compared to other MOOC predictive models that targeted different outputs from certification, such as predicting dropout rates and grades, various models were built based on public datasets (e.g. the Open University Learning Analytics Datasets [OULAD][48] and Knowledge Discovery and Datamining [KDD Cup 2015][49]). Thus, future attempts to improve or even outperform the performance of these models are initially applicable. However, this is not the case with the surveyed predictive certification models which are only applied to privacy-restricted datasets. While this might be due to the sensitivity of certification data

[48] https://analyse.kmi.open.ac.uk/open_dataset

[49] http://moocdata.cn/challenges/kdd-cup-2015

(specifically, learner financial-related data), there is no such public dataset, and many attempts to standardise MOOC datasets sharing have either met a dead end or been terminated (Lohse, McManus and Joyner, 2019).

- Regarding the "state-of-the-art" model, we note that after reviewing and synthesising the current certification predictive models in Table 3.2, it is challenging to reliably identify the state-of-the-art predictive certification model for various reasons, for example, the sources used for obtaining learners' data, the populations subset for analysis and evaluation within the surveyed works (hence the sample distributions), the algorithms and metrics used for measuring model performance, and finally, the types of the certification these works predict that are altogether different from one work to another. Thus, it is difficult, if not impossible, to determine the state-of-the-art certification predictive models. Working towards state-of-the-art models predicting MOOC certification would need an institutional initiative to provide a publicly available benchmark dataset and a protocol for executing and replicating experiments on this dataset by the researcher community.
- Finally, last but not least, the models that we have proposed here for the prediction of paid certification can be directly used by course providers, especially FutureLearn, but generalised to other platforms, to early on understand if learners are going to purchase certificates on their platform. This may enable them to early interventions, as mentioned earlier in Section 3.8, to attempt keeping those learners on the course longer, by, e.g. personalisation methods or other means. These interventions go beyond the scope of the current thesis, which proposes several ways towards predicting paid certification in Massive Open Online Courses (MOOCs). Nevertheless, this represents the most interesting avenue to explore for the future, together with exploring how our findings can be applied to other MOOC business models, as detailed in Chapter 2.

8.11. Epilogue

This chapter elucidates the outcomes of the SLR and future opportunities for development. Additionally, the statistical and ML-based results, novelty, limitations, and future works of this thesis based on the experiments conducted in Chapter 5, 6, and 7 are also discussed. The next chapter provides an overall summary of the present thesis.

Chapter 9 : Conclusion

MOOCs have gained increasing interest since their emergence due to their early commitment to openness and access to worldwide university education. Over the past decade, platforms have gathered millions of learners from different backgrounds and levels of education. Nevertheless, nowadays, the concept of MOOCs has radically changed, with the increasing emergence of MOOC monetisation (paid certification, micro-credentials, corporate training, and online degrees) and investors' willingness to return on their investments. Thus, the certification rate has been declining, due to content monetisation by platforms, resulting in a shallow rate of paid certification, where less than 5% of the total number of enrolled learners purchase a certificate at the end of their courses.

Our SLR identified the current MOOC certification predictive models (25 studies) in a systematic approach, using the PRISMA protocol. The survey concluded with an organised synthesis of the works surveyed based on their utilised input features, size, types of data used and their sources, model generalisability, prediction methods, and performance metrics reported. Additionally, the limitations of the current predictive models were identified and discussed. This includes the need for more data for a higher level of model generalisation and further validation of the achieved result, sample extensive filtration, the insufficient experiment elaboration on essential parts of experiments, the dearth of reporting on parameters fine-tuning, and non-realistic or actionable modelling. Additionally, the scarcity of early predictive models, which is essential for timely intervention, and the little models on predicting *paid* certification were also discussed.

To address these challenges, we first explored the hidden connections between learner activities and their decisions to pursue paid certification through two main approaches: (1) statistical comparisons between the activities of non-paying learners and course purchasers and (2) the application of various ML techniques to predict paid certification. Our temporal analysis conducted on a weekly basis revealed significant statistical differences in the activities between non-paying learners and certificate recipients across the five courses analysed. Additionally, we leveraged learner activities, including the number of step accesses, attempts, correct and incorrect answers, and time spent on learning steps, to develop a paid certification predictor, achieving promising BAs ranging from 0.77 to 0.95.

Subsequently, we extended our analysis to explore the wealth of information that can be extracted from MOOC interactions, particularly within discussion forums, to predict paid certification. Before delving into the discussion forums, we introduced MOOCSent, a cross-platform sentiment classifier for MOOC reviews constructed from a dataset of over 1.2 million sentiment-labeled MOOC reviews. MOOCSent addresses several limitations observed in existing sentiment classifiers, such as the reliance on single-source data and the use of two-polar classifiers (positive or negative) only. It also considers essential sentiment indicators, like emojis and emoticons, during text embedding, enabling the evaluation of model performance at a finer-grained level of sentiment.

Finally, with the assistance of MOOCSent, we employed the learners' contributions to discussion forums, to predict paid certification. This multi-input model integrated raw textual inputs from learners, sentiment classifications generated by MOOCSent, computed features (e.g., number of likes received for each textual input), and various text-related features (e.g., character counts, word counts, and part-of-speech tags for each textual instance). Our experiment incorporated diverse deep predictive approaches, particularly those supporting multi-input architectures, to conduct early weekly investigations into whether data derived from MOOC learners' interactions in discussion forums can anticipate learners' decisions to pursue paid certification.

In conclusion, this thesis makes concrete contributions to the field of MOOC learner analytics by (1) exploring various conventional and deep ML methods for predicting paid certification in MOOCs based on learner clickstream data and course discussion forums, (2) developing MOOCSent, the most extensive MOOC sentiment classifier, by analysing learners' reviews from leading MOOC platforms, namely Coursera, FutureLearn, and Udemy, and incorporating specialised lexicons encompassing over three thousand corresponding explanatory words and phrases related to emojis and emoticons, and (3) introducing an innovative multi-input model for predicting certification based on data from discussion forums, which simultaneously processes textual elements (comments and replies) and numerical data (e.g., number of likes posted and received, sentiments).

Appendix A

A list of MOOC Platforms distributed by the country of establishment, own elaboration.

Country	Platforms	Total number
United States	Coursera, edX, Udemy, Udacity, Canvas, MIT OpenCourseWare, Khan, LinkedIn, Kadenze, skillShare, Domestika, CreativeLive, Microsoft Learn, The Great Course, HubSpot, Brilliant, Stanford Languita	17
China	XuetangX, CNMOOC, Chinese University MOOC, Zhihuishu, Fanya, Xue Yin	6
Japan	Schoo, gacco, Fisdom, JMOOC, OpenLearning	5
Italy	Federica, Polimi, EduOpen, EMMA	4
France	OpenClassrooms, FUN, OpenSAP	3
India	Swayam, Edureka, NPTEL	3
Saudi Arabia	kkuX, Rawaq, Future X	3
Ireland	Alison, Shaw Academy	2
Taiwan	Open Education, ewant	2
Australia	OpenLearning, openUniversity	2
Germany	iversity, openHPI	2
Israel	Campus-IL	1
Jordan	Edraak	1
United Kingdom	FutureLearn	1
Indonesia	IndonesiaX	1
Finland	MOOC.fi	1
Spain	Miriada X	1
Mexico	MéxicoX	1
Russia	Open Education	1
Ukraine	Prometheus	1

Thailand	ThaiMOOC	1
Brazil	Veduca	1
South Korea	k-MOOC	1
Austria	iMooX	1
Belgium	KU Leuven	1
Total		63

Appendix B

Penn POS tags

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Appendix C

PRISMA 2020 Checklist

Section and Topic	Checklist item	Page Number
TITLE		
Title	Identify the report as a systematic review.	29
INTRODUCTION		
Rationale	Describe the rationale for the review in the context of existing knowledge.	30
Objectives	Provide an explicit statement of the objective(s) or question(s) the review addresses.	29
METHODS		
Eligibility criteria	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	31
Information sources	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	31
Search strategy	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	40
Selection process	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	35
Data collection process	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	37

Data items	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	51
	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	51
Study risk of bias assessment	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	n/a
Effect measures	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	n/a
Synthesis methods	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	n/a
	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	n/a
	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	39
	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	n/a
	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	n/a
	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	n/a
Reporting bias assessment	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	n/a
Certainty assessment	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	n/a
RESULTS		
Study selection	Describe the results of the search and selection process, from the number of records identified in the search to the number	39

	of studies included in the review, ideally using a flow diagram.	
	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	37
Study characteristics	Cite each included study and present its characteristics.	40
Risk of bias in studies	Present assessments of risk of bias for each included study.	n/a
Results of individual studies	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	n/a
Results of syntheses	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	65
	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	65
	Present results of all investigations of possible causes of heterogeneity among study results.	n/a
	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	n/a
Reporting biases	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	n/a
Certainty of evidence	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	n/a
DISCUSSION		
Discussion	Provide a general interpretation of the results in the context of other evidence.	51
	Discuss any limitations of the evidence included in the review.	69
	Discuss any limitations of the review processes used.	31
	Discuss implications of the results for practice, policy, and future research.	69
OTHER INFORMATION		
Registration and protocol	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	n/a

	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	n/a
	Describe and explain any amendments to information provided at registration or in the protocol.	n/a
Support	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	n/a
Competing interests	Declare any competing interests of review authors.	n/a
Availability of data, code and other materials	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	39

References

ACM Advanced Search. Available at: https://dl.acm.org/search/advanced (Accessed: 24/11/2021 2021).

Agrawal, A., Venkatraman, J., Leonard, S. and Paepcke, A. (2015) 'YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips'.

Ahuja, S. and Dubey, G. 'Clustering and sentiment analysis on Twitter data'. *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*: IEEE, 1-5.

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. 'Optuna: A next-generation hyperparameter optimization framework'. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623-2631.

Ali, J., Khan, R., Ahmad, N. and Maqsood, I. (2012) 'Random forests and decision trees', *International Journal of Computer Science Issues (IJCSI),* 9(5), pp. 272.

Aljohani, T. and Cristea, A. I. 'Training Temporal and NLP Features via Extremely Randomised Trees for Educational Level Classification'. *International Conference on Intelligent Tutoring Systems*: Springer, 136-147.

Almatrafi, O. and Johri, A. (2018) 'Systematic review of discussion forums in massive open online courses (MOOCs)', *IEEE Transactions on Learning Technologies,* 12(3), pp. 413-428.

Alpaydin, E. (2020) *Introduction to machine learning.* MIT press.

Alrajhi, L., Pereira, F. D., Cristea, A. I. and Aljohani, T. 'A Good Classifier is Not Enough: A XAI Approach for Urgent Instructor-Intervention Models in MOOCs'. *International Conference on Artificial Intelligence in Education*: Springer, 424-427.

Alshehri, M., Alamri, A. and Cristea, A. I. 'Predicting Certification in MOOCs Based on Students' Weekly Activities'. *International Conference on Intelligent Tutoring Systems*: Springer, 173-185.

Alshehri, M., Alamri, A., Cristea, A. I. and Stewart, C. D. (2021) 'Towards Designing Profitable Courses: Predicting Student Purchasing Behaviour in MOOCs', *International Journal of Artificial Intelligence in Education*, pp. 1-19.

Alshehri, M., Foss, J., Cristea, A. I., Kayama, M., Shi, L., Alamri, A. and Tsakalidis, A. 'On the need for fine-grained analysis of Gender versus Commenting Behaviour in MOOCs'. *Proceedings of the 2018 The 3rd International Conference on Information and Education Innovations*: ACM, 73-77.

Alsheri, M. A., Almari, A., Cristea, A. I. and Stewart, C. D. 'forum-based Prediction of Certification in Massive Open Online Courses'. Association for Information Systems.

Arslan, F., Bagchi, K. and Ryu, S. (2015) 'A Preliminary Evaluation of the determinants of certification success in MOOCs: A multi-level study'.

Asharf, J., Moustafa, N., Khurshid, H., Debie, E., Haider, W. and Wahab, A. (2020) 'A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions', *Electronics,* 9(7), pp. 1177.

Atiaja, L. A. and Proenza, R. (2016) 'The MOOCs: origin, characterization, principal problems and challenges in Higher Education', *Journal of e-learning and Knowledge Society,* 12(1).

Badali, M., Hatami, J., Banihashem, S. K., Rahimi, E., Noroozi, O. and Eslami, Z. (2022) 'The role of motivation in MOOCs' retention rates: a systematic literature review', *Research and Practice in Technology Enhanced Learning,* 17(1), pp. 1-20.

Baker, R. M. and Passmore, D. L. (2016) 'Value and pricing of MOOCs', *Education Sciences,* 6(2), pp. 14.

Baker, R. S. and Hawn, A. (2021) 'Algorithmic bias in education', *International Journal of Artificial Intelligence in Education,* pp. 1-41.

Bakharia, A. 'Towards cross-domain MOOC forum post classification'. *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, 253-256.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E. and Acharya, U. R. (2021) 'ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis', *Future Generation Computer Systems,* 115, pp. 279-294.

Baturay, M. H. (2015) 'An overview of the world of MOOCs', *Procedia-Social and Behavioral Sciences,* 174, pp. 427-433.

Bayeck, R. (2016) 'Exploratory study of MOOC learners' demographics and motivation: The case of students involved in groups', *Open Praxis,* 8(3), pp. 223-233.

Bayer, M., Kaufhold, M.-A. and Reuter, C. (2022) 'A survey on data augmentation for text classification', *ACM Computing Surveys,* 55(7), pp. 1-39.

Beckerle, M., Chatzopoulou, A. and Fischer-Hübner, S. 'Towards Cybersecurity MOOC Certification'. *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*: IEEE, 1-11.

Belleflamme, P. and Jacqmin, J. (2016) 'An economic appraisal of MOOC platforms: Business models and impacts on higher education', *CESifo Economic Studies,* 62(1), pp. 148-169.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D. and Cox, D. D. (2015) 'Hyperopt: a python library for model selection and hyperparameter optimization', *Computational Science & Discovery,* 8(1), pp. 014008.

Bhardwaj, A., Di, W. and Wei, J. (2018) *Deep Learning Essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling.* Packt Publishing Ltd.

Blackmon, S. J. and Major, C. H. (2016) *MOOCs and Higher Education: Implications for Institutional Research: New Directions for Institutional Research, Number 167.* John Wiley & Sons.

Bogdan, R., Holotescu, C., Andone, D. and Grosseck, G. (2017) 'How MOOCs are being used for corporate training?', *eLearning & Software for Education*, 2.

Bonafini, F. C. (2017) 'The effects of participants' engagement with videos and forums in a MOOC for teachers' professional development', *Open Praxis*, 9(4), pp. 433-447.

Bonafini, F. C. (2018) 'Characterizing Super-Posters in a MOOC for Teachers' Professional Development', *Online Learning*, 22(4), pp. 89-108.

Bonta, V. and Janardhan, N. K. a. N. (2019) 'A comprehensive study on lexicon based approaches for sentiment analysis', *Asian Journal of Computer Science and Technology*, 8(S2), pp. 1-6.

Borrás-Gené, O. 'Empowering MOOC participants: Dynamic content adaptation through external tools'. *European MOOCs Stakeholders Summit*: Springer, 121-130.

Borrego, Á. (2019) 'The impact of MOOCs on library and information science education', *Education for Information*, 35(2), pp. 87-98.

Breiman, L. (2001) 'Random forests', *Machine learning*, 45(1), pp. 5-32.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D. and Seaton, D. T. (2013) 'Studying learning in the worldwide classroom research into edX's first MOOC', *Research & Practice in Assessment*, 8, pp. 13-25.

Briggs, J. (2021) 'Masked-Language Modeling With BERT'. Available at: https://towardsdatascience.com/masked-language-modelling-with-bert-7d49793e5d2c (Accessed 10/11/2022).

Brouns, F., Mota, J., Morgado, L., Jansen, D., Fano, S., Silva, A. and Teixeira, A. M. 'A networked learning framework for effective MOOC design: the ECO project approach'. *EDEN Conference Proceedings*, 161-172.

Brown, S. 'Back to the future with MOOCs'. *ICICTE 2013 Proceedings*, 237-246.

Brownlee, J. (2021) 'A Gentle Introduction to Ensemble Learning Algorithms', *Machine Learning Mastery*.

Buholzer, F., Rietsche, R. and Söllner, M. (2018) 'Knowing what learners like– Developing a cultural sensitive peer assessment process in MOOCs'.

Bulusu, A. and Rao, K. R. 'Sentiment analysis of learner reviews to improve efficacy of massive open online courses (MOOC’s)-A survey'. *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*: IEEE, 933-941.

Burd, E. L., Smith, S. P. and Reisman, S. (2015) 'Exploring business models for MOOCs in higher education', *Innovative Higher Education,* 40(1), pp. 37-49.

Cagiltay, N. E., Cagiltay, K. and Celik, B. (2020) 'An Analysis of Course Characteristics, Learner Characteristics, and Certification Rates in MITx MOOCs', *International Review of Research in Open and Distributed Learning,* 21(3), pp. 121-139.

Calise, M., Kloos, C. D., Reich, J., Ruiperez-Valiente, J. A. and Wirsing, M. (2019) *Digital Education: At the MOOC Crossroads Where the Interests of Academia and Business Converge: 6th European MOOCs Stakeholders Summit, EMOOCs 2019, Naples, Italy, May 20–22, 2019, Proceedings.* Springer.

Canessa, E., Tenze, L. and Salvatori, E. (2013) 'Attendance to massive open on-line courses: Towards a solution to track on-line recorded lectures viewing', *Bulletin of the IEEE Technical Committee on Learning Technology,* 15(1), pp. 36-39.

Casserly, C. (2018) '10 years of OER: What funders can learn from a historical moment', *Hewlett Foundation*.

Castillo, N. M., Lee, J., Zahra, F. T. and Wagner, D. A. (2015) 'MOOCS for development: Trends, challenges, and opportunities', *International Technologies & International Development,* 11(2), pp. 35.

Celik, B. and Cagiltay, K. (2023) 'Did you act according to your intention? An analysis and exploration of intention–behavior gap in MOOCs', *Education and Information Technologies*, pp. 1-28.

Chaplot, D. S., Rhim, E. and Kim, J. 'Predicting student attrition in MOOCs using sentiment analysis and neural networks'. *Work. 17th Int. Conf. Artif. Intell. Educ. AIED-WS 2015*, 7-12.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of artificial intelligence research,* 16, pp. 321-357.

Chawla, S. (2021) 'Application of convolution neural network in web query session mining for personalised web search', *International Journal of Computational Science and Engineering,* 24(4), pp. 417-428.

Chen, J., Feng, J., Sun, X. and Liu, Y. (2019) 'Co-training semi-supervised deep learning for sentiment classification of MOOC forum posts', *Symmetry,* 12(1), pp. 8.

Chen, T. and Guestrin, C. 'Xgboost: A scalable tree boosting system'. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.

Chiorrini, A., Diamantini, C., Mircoli, A. and Potena, D. 'Emotion and sentiment analysis of tweets using BERT'. *EDBT/ICDT Workshops*.

Cho, K., Van Merriënboer, B., Bahdanau, D. and Bengio, Y. (2014) 'On the properties of neural machine translation: Encoder-decoder approaches', *arXiv preprint arXiv:1409.1259*.

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D. and Emanuel, E. (2013) 'The MOOC phenomenon: Who takes massive open online courses and why?', *Available at SSRN 2350964*.

Christie, S. T., Jarratt, D. C., Olson, L. A. and Taijala, T. T. (2019) 'Machine-Learned School Dropout Early Warning at Scale', *International Educational Data Mining Society*.

Chuang, I. and Ho, A. (2016) 'HarvardX and MITx: Four Years of Open Online Courses--Fall 2012-Summer 2016'.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014) 'Empirical evaluation of gated recurrent neural networks on sequence modeling', *arXiv preprint arXiv:1412.3555*.

Clavié, B. and Gal, K. (2019) 'Edubert: Pretrained deep language models for learning analytics', *arXiv preprint arXiv:1912.00690*.

Cliche, M. (2017) 'BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs', *arXiv preprint arXiv:1704.06125*.

Clow, D. 'MOOCs and the funnel of participation'. *Proceedings of the third international conference on learning analytics and knowledge*: ACM, 185-189.

Cobos, R. and Jurado, F. 'An exploratory analysis on MOOCs retention and certification in two courses of different knowledge areas'. *2018 IEEE Global Engineering Education Conference (EDUCON)*: IEEE, 1659-1666.

Cobos, R. and Olmos, L. 'A learning analytics tool for predictive modeling of dropout and certificate acquisition on MOOCs for professional learning'. *2018 IEEE international conference on industrial engineering and engineering management (IEEM)*: IEEE, 1533-1537.

Cohen, A., Shimony, U., Nachmias, R. and Soffer, T. (2019) 'Active learners’ characterization in MOOC forums and their generated knowledge', *British journal of educational technology*, 50(1), pp. 177-198.

Coleman, C. A., Seaton, D. T. and Chuang, I. 'Probabilistic use cases: Discovering behavioral patterns for predicting certification'. *Proceedings of the second (2015) acm conference on learning@ scale*, 141-148.

Condé, J. and Cisel, M. 'On the use of MOOCs in companies: A panorama of current practices'. *European MOOCs Stakeholders Summit*: Springer, 37-46.

Cramer, J. S. (2002) 'The origins of logistic regression'.

Cusumano, M. A. (2013) 'Are the costs of'free'too high in online education?', *Communications of the ACM,* 56(4), pp. 26-28.

Dalipi, F., Imran, A. S. and Kastrati, Z. 'MOOC dropout prediction using machine learning techniques: Review and research challenges'. *Global Engineering Education Conference (EDUCON), 2018 IEEE*: IEEE, 1007-1014.

Dalipi, F., Zdravkova, K. and Ahlgren, F. (2021) 'Sentiment analysis of students' feedback in MOOCs: A systematic literature review', *Frontiers in Artificial Intelligence,* 4, pp. 728708.

Daniel, J. (2012) 'Making sense of MOOCs: Musings in a maze of myth, paradox and possibility', *Journal of interactive Media in education,* 2012(3).

Davis, H., Leon, K. D. M., Vera, M. and White, S. (2013) 'MOOCs for Universities and Learners', *An analysis of motivating factors*.

DeBoer, J., Ho, A. D., Stump, G. S. and Breslow, L. (2014) 'Changing "course" reconceptualizing educational variables for massive open online courses', *Educational researcher,* 43(2), pp. 74-84.

DeBoer, J., Stump, G. S., Seaton, D., Ho, A., Pritchard, D. E. and Breslow, L. 'Bringing student backgrounds online: MOOC user demographics, site usage, and online learning'. *Educational data mining 2013*.

Dellarocas, C. and Van Alstyne, M. W. (2013) 'Money models for MOOCs', *Communications of the ACM, August,* 56(8), pp. 25-28.

Deng, L. and Yu, D. (2014) 'Deep learning: methods and applications', *Foundations and trends® in signal processing,* 7(3–4), pp. 197-387.

Deng, R., Benckendorff, P. and Gannaway, D. (2019) 'Progress and new directions for teaching and learning in MOOCs', *Computers & Education,* 129, pp. 48-60.

developers, s.-l. (2007-2020) *Metrics and scoring: quantifying the quality of predictions*. Available at: https://scikit-learn.org/stable/modules/model_evaluation.html#balanced-accuracy-score (Accessed: 30/03/2021 2021).

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*.

Dillahunt, T., Wang, Z. and 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), pp. 177-196.

Diver, P. and Martinez, I. (2015) 'MOOCs as a massive research laboratory: Opportunities and challenges', *Distance Education,* 36(1), pp. 5-25.

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G. and Rangwala, H. (2016) 'Predicting student performance using personalized analytics', *Computer,* 49(4), pp. 61-69.

Farrow, R. (2019) 'Massive Open Online Courses for Business Learning: Key research, best practices and pathways to innovation'.

Fedorova, E. P. and Skobleva, E. I. (2020) 'Application of blockchain technology in higher education', *European Journal of Contemporary Education,* 9(3), pp. 552-571.

Ferguson, R. and Clow, D. 'Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs)'. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*: ACM, 51-58.

Fini, A. (2009) 'The technological dimension of a massive open online course: The case of the CCK08 course tools', *The International Review of Research in Open and Distributed Learning,* 10(5).

Fleiss, J. L., Levin, B. and Paik, M. C. (1981) 'The measurement of interrater agreement', *Statistical methods for rates and proportions,* 2(212-236), pp. 22-23.

Fleming, P. S., Koletsi, D. and Pandis, N. (2014) 'Blinded by PRISMA: are systematic reviewers focusing on PRISMA and ignoring other guidelines?', *PLoS One,* 9(5), pp. e96407.

Fotso, J. E. M., Batchakui, B., Nkambou, R. and Okereke, G. (2022) 'Algorithms for the development of deep learning models for classification and prediction of learner behaviour in moocs', *Artificial Intelligence for Data Science in Theory and Practice*: Springer, pp. 41-73.

Freund, Y. and Schapire, R. E. (1997) 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of computer and system sciences,* 55(1), pp. 119-139.

Friedman, J. H. (2001) 'Greedy function approximation: a gradient boosting machine', *Annals of statistics,* pp. 1189-1232.

Garcia Barrera, A., Gomez Hernandez, P. and Monge Lopez, C. (2017) 'ATTENTION TO DIVERSITY IN MOOCS: A METHODOLOGICAL PROPOSAL', *EDUCACION XX1,* 20(2), pp. 215-233.

García-Molina, S., Alario-Hoyos, C., Moreno-Marcos, P. M., Muñoz-Merino, P. J., Estévez-Ayres, I. and Delgado Kloos, C. (2020) 'An algorithm and a tool for the automatic grading of MOOC learners from their contributions in the discussion forum', *Applied Sciences,* 11(1), pp. 95.

Gardair, C., Bousquet, G., Lehmann-Che, J., de Bazelaire, C., de Cremoux, P., Van Nhieu, J. T., Sockeel, M., Battistella, M., Calvani, J. and Gervais, J. 'Les coulisses d’un Massive Open Online Course (MOOC) sur le diagnostic des cancers'. *Annales de Pathologie*: Elsevier, 305-311.

Gardner, J. and Brooks, C. 'Dropout model evaluation in MOOCs'. *Thirty-Second AAAI Conference on Artificial Intelligence*.

Gardner, J. and Brooks, C. (2018b) 'Student success prediction in MOOCs', *User Modeling and User-Adapted Interaction*, 28(2), pp. 127-203.

Gardner, J., Brooks, C. and Baker, R. 'Evaluating the fairness of predictive student models through slicing analysis'. *Proceedings of the 9th international conference on learning analytics & knowledge*, 225-234.

Geurts, P., Ernst, D. and Wehenkel, L. (2006) 'Extremely randomized trees', *Machine learning*, 63(1), pp. 3-42.

Gitinabard, N., Khoshnevisan, F., Lynch, C. F. and Wang, E. Y. (2018) 'Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features', *arXiv preprint arXiv:1809.00052*.

Glance, D. G., Barrett, P. H. R. and Hugh, R. 'Attrition patterns amongst participant groups in Massive Open Online Courses'. *ASCILITE Conference, Dunedin, New Zealand. Retrieved from http://ascilite2014. otago. ac. nz/files/fullpapers/16-Glance. pdf.*

Glass, C. R., Shiokawa-Baklan, M. S. and Saltarelli, A. J. (2016) 'Who takes MOOCs?', *New Directions for Institutional Research*, 2015(167), pp. 41-55.

Goli, A., Chintagunta, P. K. and Sriram, S. (2019) 'Effect of Payment on User Engagement in MOOCs', *Available at SSRN 3414406*.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J. and Sculley, D. 'Google vizier: A service for black-box optimization'. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1487-1495.

González Robinson, K. (2016) 'New internationalization opportunities for Higher Education Institutions: A strategic framework for the cross-border provision of Massive Open Online Courses (MOOCs)'.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep learning.* MIT press.

Greene, J. A., Oswald, C. A. and Pomerantz, J. (2015) 'Predictors of retention and achievement in a massive open online course', *American Educational Research Journal*, 52(5), pp. 925-955.

Haddadi, L. and Dahmani, F. B. 'An assessment planner for MOOCs based ODALA approach'. *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*: IEEE, 855-862.

Halsbenning, S. and Niemann, M. 'Sustainable MOOC Platforms-Searching for Business Models of the Future'. *ECIS*.

He, J., Bailey, J., Rubinstein, B. and Zhang, R. 'Identifying at-risk students in massive open online courses'. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Henukh, A., Nikat, R., Simbolon, M., Nuryadin, C. and Baso, Y. 'Multimedia development based on web connected Massive Open Online Courses (cMOOCs) on the basic physics material'. *IOP Conference Series: Earth and Environmental Science*: IOP Publishing, 012160.

Hill, P. (2012) 'Four Barriers that MOOCs must overcome to build a sustainable model', *Recuperado el*, 1, pp. 166.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural computation*, 9(8), pp. 1735-1780.

Hollands, F. and Kazi, A. (2019) 'MOOC-based alternative credentials: What's the value for the learner', *Educause Review*.

Hone, K. S. and El Said, G. R. (2016) 'Exploring the factors affecting MOOC retention: A survey study', *Computers & Education*, 98, pp. 157-168.

Hutter, F., Hoos, H. H. and Leyton-Brown, K. 'Sequential model-based optimization for general algorithm configuration'. *International conference on learning and intelligent optimization*: Springer, 507-523.

Hutto, C. and Gilbert, E. 'Vader: A parsimonious rule-based model for sentiment analysis of social media text'. *Proceedings of the international AAAI conference on web and social media*, 216-225.

IEEE Explore Search Tips. Available at: https://ieeexplore.ieee.org/Xplorehelp/searching-ieee-xplore/search-tips (Accessed: 24/11/2021 2021).

Iiyoshi, T. and Kumar, M. (2010) *Opening up education: The collective advancement of education through open technology, open content, and open knowledge.* The MIT Press.

Ilavarasan, E. 'A Survey on Sarcasm detection and challenges'. *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*: IEEE, 1234-1240.

Impey, C. D., Wenger, M. C. and Austin, C. L. (2015) 'Astronomy for astronomical numbers: A worldwide massive open online class', *International Review of Research in Open and Distributed Learning*, 16(1), pp. 57-79.

Isidro, C., Carro, R. M. and Ortigosa, A. 'Dropout detection in MOOCs: An exploratory analysis'. *2018 International Symposium on Computers in Education (SIIE)*: IEEE, 1-6.

Jaganathan, G., Sugundan, N. and Sivakumar, S. (2018) 'MOOCs: A Comparative analysis between Indian scenario and Global scenario', *International Journal of Engineering & Technology*, 7(4), pp. 854-857.

Jahan, M. S., Beddiar, D. R., Oussalah, M. and Arhab, N. (2021) 'Hate and Offensive language detection using BERT for English Subtask A'.

Jena, R. (2018) 'Predicting students’ learning style using learning analytics: a case study of business management students from India', *Behaviour & Information Technology,* 37(10-11), pp. 978-992.

Jiang, S., Fitzhugh, S. M. and Warschauer, M. 'Social positioning and performance in MOOCs'. *Workshop on graph-based educational data mining*.

Jiang, S., Williams, A., Schenke, K., Warschauer, M. and O'dowd, D. 'Predicting MOOC performance with week 1 behavior'. *Educational data mining 2014*.

Jiang, Z., Zhang, Y. and Li, X. (2015) 'Learning behavior analysis and prediction based on MOOC data', *Journal of computer research and development,* 52(3), pp. 614.

Joksimović, S., Manataki, A., Gašević, D., Dawson, S., Kovanović, V. and De Kereki, I. F. (2016) 'Translating network position into performance: Importance of centrality in different network configurations'. *Proceedings of the sixth international conference on learning analytics & knowledge*, 314-323.

Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., Dawson, S., Graesser, A. C. and Brooks, C. (2018) 'How do we model learning at scale? A systematic review of research on MOOCs', *Review of Educational Research,* 88(1), pp. 43-86.

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), pp. 133-160.

Jose, K. (2020) 'RNNs, LSTMs, CNNs, Transformers and BERT'. Available at: https://medium.com/analytics-vidhya/rnns-lstms-cnns-transformers-and-bert-be003df3492b.

Kastrati, Z., Arifaj, B., Lubishtani, A., Gashi, F. and Nishliu, E. 'Aspect-Based Opinion Mining of Students' Reviews on Online Courses'. *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, 510-514.

Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K. and Wani, M. A. (2021) 'Sentiment analysis of students’ feedback with NLP and deep learning: A systematic mapping study', *Applied Sciences,* 11(9), pp. 3986.

Kastrati, Z., Imran, A. S. and Kurti, A. (2020) 'Weakly supervised framework for aspect-based sentiment analysis on students’ reviews of MOOCs', *IEEE Access,* 8, pp. 106799-106810.

Khan, M. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S. and Raazi, S. M. K.-u.-R. (2021a) 'Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques', *Complexity,* 2021.

Khan, P., Kader, M. F., Islam, S. R., Rahman, A. B., Kamal, M. S., Toha, M. U. and Kwak, K.-S. (2021b) 'Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances', *IEEE Access*, 9, pp. 37622-37655.

King, I. and Lee, W.-I. (2022) 'Exploring Global MOOC Ecosystems', *A Decade of MOOCs and Beyond: Platforms, Policies, Pedagogy, Technology, and Ecosystems with an Emphasis on Greater China*: Springer, pp. 117-132.

Kite, J., Indig, D., Mihrshahi, S., Milat, A. and Bauman, A. (2015) 'Assessing the usefulness of systematic reviews for policymakers in public health: a case study of overweight and obesity prevention interventions', *Preventive Medicine*, 81, pp. 99-107.

Kizilcec, R. F. and Halawa, S. 'Attrition and achievement gaps in online learning'. *Proceedings of the second (2015) ACM conference on learning@ scale*, 57-66.

Kizilcec, R. F., Piech, C. and Schneider, E. 'Deconstructing disengagement: analyzing learner subpopulations in massive open online courses'. *Proceedings of the third international conference on learning analytics and knowledge*: ACM, 170-179.

Kocdar, S., OKUR, M. R. and Bozkurt, A. (2017) 'An examination of xMOOCS: An embedded single case study based on Conole’s 12 dimensions', *Turkish Online Journal of Distance Education*, 18(4), pp. 52-65.

Koch, P., Golovidov, O., Gardner, S., Wujek, B., Griffin, J. and Xu, Y. 'Autotune: A derivative-free optimization framework for hyperparameter tuning'. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 443-452.

Koller, D., Ng, A., Do, C. and Chen, Z. (2013) 'Retention and intention in massive open online courses: In depth', *Educause review*, 48(3), pp. 62-63.

Kostopoulos, G., Panagiotakopoulos, T., Kotsiantis, S., Pierrakeas, C. and Kameas, A. (2021) 'Interpretable Models for Early Prediction of Certification in MOOCs: A Case Study on a MOOC for Smart City Professionals', *IEEE Access*, 9, pp. 165881-165891.

Kour, H. and Gupta, M. K. (2022) 'An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM', *Multimedia Tools and Applications*, pp. 1-37.

Koutropoulos, A. (2013) 'MOOCs in Higher Education: Options, Affordances, Pitfalls (Part 2)', *Learning Solutions Magazine*.

Krauss, S. M. (2017) 'How competency-based education may help reduce our nation’s toughest inequities', *Lumina Issue Papers. Lumina Foundation. http://hdl. handle. net/10919/83258*.

Kučak, D., Juričić, V. and Đambić, G. (2018) 'MACHINE LEARNING IN EDUCATION-A SURVEY OF CURRENT RESEARCH TRENDS', *Annals of DAAAM & Proceedings*, 29.

Kuhn, M. and Johnson, K. (2013) *Applied predictive modeling.* Springer.

Kumar, K. (2019) 'A study of veterinary scholars’ perception of MOOCs', *Information and Learning Sciences*.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *nature,* 521(7553), pp. 436-444.

Lee, M. K. (2018a) 'Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management', *Big Data & Society,* 5(1), pp. 2053951718756684.

Lee, Y. (2018b) 'Effect of uninterrupted time-on-task on students’ success in Massive Open Online Courses (MOOCs)', *Computers in Human Behavior,* 86, pp. 174-180.

Lee, Y. (2019) 'Using self-organizing map and clustering to investigate problem-solving patterns in the massive open online course: An exploratory study', *Journal of Educational Computing Research,* 57(2), pp. 471-490.

Lemoine, P. A. and Richardson, M. D. (2015) 'Micro-credentials, nano degrees, and digital badges: New credentials for global higher education', *International Journal of Technology and Educational Marketing (IJTEM),* 5(1), pp. 36-49.

Lewis, M. J. and Lodge, J. M. (2016) 'Keep calm and credential on: Linking learning, life and work practices in a complex world', *Foundation of digital badges and micro-credentials*: Springer, pp. 41-54.

Li, X., Bing, L., Zhang, W. and Lam, W. (2019a) 'Exploiting BERT for end-to-end aspect-based sentiment analysis', *arXiv preprint arXiv:1910.00883*.

Li, X., Zhang, H., Ouyang, Y., Zhang, X. and Rong, W. 'A shallow BERT-CNN model for sentiment analysis on MOOCs comments'. *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*: IEEE, 1-6.

Liao, P., Sun, Y., Ye, S., Li, X., Su, G. and Sun, Y. 'Predicting learners' multi-question performance based on neural networks'. *2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*: IEEE, 1-6.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J. and Moher, D. (2009) 'The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration', *Journal of clinical epidemiology,* 62(10), pp. e1-e34.

Lim, C. L., Tang, S. F. and Ravichandran, P. 'A Study on the Mediation Effects of Intention to Enroll in MOOCs on its Actual Usage'. *Proceedings of the 8th International Conference on E-Education, E-Business, E-Management and E-Learning*, 30-33.

Littenberg-Tobias, J. and Reich, J. (2020) 'Evaluating access, quality, and equity in online learning: A case study of a MOOC-based blended professional degree program', *The Internet and Higher Education*, 47, pp. 100759.

Littenberg-Tobias, J., Ruipérez-Valiente, J. A. and Reich, J. (2020) 'Studying learner behavior in online courses with free-certificate coupons: Results from two case studies', *International Review of Research in Open and Distributed Learning*, 21(1), pp. 1-22.

Liu, S., Liu, S., Liu, Z., Peng, X. and Yang, Z. (2022a) 'Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement', *Computers & Education*, 181, pp. 104461.

Liu, Z., Mu, R., Yang, Z., Peng, X., Liu, S. and Chen, J. (2022b) 'Modeling temporal cognitive topic to uncover learners' concerns under different cognitive engagement patterns', *Interactive Learning Environments*, pp. 1-18.

Liyanagunawardena, T. R., Lundqvist, K., Mitchell, R., Warburton, S. and Williams, S. A. (2019) 'A MOOC Taxonomy Based on Classification Schemes of MOOCs', *European Journal of Open, Distance and E-learning*, 22(1), pp. 85-103.

Lohr, S. (2020) 'Remember the MOOCs? After near-death, they're booming', *The New York Times*, 26.

Lohse, J. J., McManus, C. A. and Joyner, D. A. 'Surveying the MOOC data set universe'. *2019 IEEE Learning With MOOCS (LWMOOCS)*: IEEE, 159-164.

Longstaff, E. (2014) 'The prehistory of MOOCs: Inclusive and exclusive access in the cyclical evolution of Higher Education', *Journal of Organisational Transformation & Social Change*, 11(3), pp. 164-184.

lopez, d. (2019) *Recurrent Neural Network (RNN), Long-Short Term Memory (LSTM) & Gated Recurrent Unit (GRU)*. Available at: http://dprogrammer.org/rnn-lstm-gru (Accessed: 22/10/2022.

Loria, S. (2018) 'textblob Documentation', *Release 0.15*, 2.

Lundqvist, K., Liyanagunawardena, T. and Starkey, L. (2020) 'Evaluation of student feedback within a MOOC using sentiment analysis and target groups', *International Review of Research in Open and Distributed Learning*, 21(3), pp. 140-156.

Macleod, H., Haywood, J., Woodgate, A. and Alkhatnai, M. (2015) 'Emerging patterns in MOOCs: Learners, course designs and directions', *TechTrends*, 59(1), pp. 56-63.

Makel, M. C. and Plucker, J. A. (2014) 'Facts are more important than novelty: Replication in the education sciences', *Educational Researcher*, 43(6), pp. 304-316.

Malko, A., Paris, C., Duenser, A., Kangas, M., Molla, D., Sparks, R. and Wan, S. 'Demonstrating the reliability of self-annotated emotion data'. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 45-54.

Marcus, M., Santorini, B. and Marcinkiewicz, M. A. (1993) 'Building a large annotated corpus of English: The Penn Treebank'.

McGreal, R., Kinuthia, W., Marshall, S. and McNamara, T. (2013) *Open educational resources: Innovation, research and practice.* Commonwealth of Learning.

McHugh, M. L. (2012) 'Interrater reliability: the kappa statistic', *Biochemia medica,* 22(3), pp. 276-282.

McKnight, P. E. and Najab, J. (2010) 'Mann‐Whitney U Test', *The Corsini encyclopedia of psychology*, pp. 1-1.

Medhat, W., Hassan, A. and Korashy, H. (2014) 'Sentiment analysis algorithms and applications: A survey', *Ain Shams engineering journal,* 5(4), pp. 1093-1113.

Mehrabi, M., Safarpour, A. R. and Keshtkar, A. A. (2020) 'Massive Open Online Courses (MOOCs) dropout rate in the world: A systematic review protocol'.

Michael Spector, J. (2017) 'A critical look at MOOCs', *Open education: From OERs to MOOCs*: Springer, pp. 135-147.

Milheim, W. D. (2013) 'Massive open online courses (MOOCs): Current applications and future potential', *Educational Technology*, pp. 38-42.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. (1990) 'Introduction to WordNet: An on-line lexical database', *International journal of lexicography,* 3(4), pp. 235-244.

Milligan, C. and Littlejohn, A. (2014) 'Supporting professional learning in a massive open online course', *International Review of Research in Open and Distributed Learning,* 15(5), pp. 197-213.

Mirza, M., Lukosch, S. and Lukosch, H. 'Twitter Sentiment Analysis of Cross-Cultural Perspectives on Climate Change'. *International Conference on Human-Computer Interaction*: Springer, 392-406.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. and Group*, P. (2009) 'Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement', *Annals of internal medicine,* 151(4), pp. 264-269.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., Estévez-Ayres, I. and Kloos, C. D. 'Sentiment analysis in MOOCs: A case study'. *2018 IEEE Global Engineering Education Conference (EDUCON)*: IEEE, 1489-1496.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J. and Kloos, C. D. (2018b) 'Prediction in MOOCs: A review and future research directions', *IEEE Transactions on Learning Technologies,* 12(3), pp. 384-401.

Mourdi, Y., Sadgal, M., El Kabtane, H. and Fathi, W. B. (2019) 'A machine learning-based methodology to predict learners’ dropout, success or failure in MOOCs', *International Journal of Web Information Systems*.

Mullaney, T. and Reich, J. 'Staggered versus all-at-once content release in massive open online courses: Evaluating a natural experiment'. *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 185-194.

Munigadiapa, P. and Adilakshmi, T. (2022) 'MOOC-LSTM: The LSTM Architecture for Sentiment Analysis on MOOCs Forum Posts', *Computational Intelligence and Data Analytics: Proceedings of ICCIDA 2022*: Springer, pp. 283-293.

Ng, A. and Widom, J. (2014) 'Origins of the Modern MOOC (xMOOC)', *Hrsg. Fiona M. Hollands, Devayani Tirthali: MOOCs: Expectations and Reality: Full Report*, pp. 34-47.

Njingang Mbadjoin, T. and Chaker, R. (2021) 'Les liens entre les objectifs de formation, les facteurs sociodemographiques et la reussite chez des participants a un MOOC professionnalisant', *McGill Journal of Education/Revue des sciences de l'éducation de McGill,* 56(1), pp. 149-170.

Nkuyubwatsi, B. (2014) 'Cultural translation in massive open online courses (MOOCs)', *EMOOCs*, pp. 122-129.

Notaris, D. D. 'Reskilling Higher Education Professionals'. *European MOOCs Stakeholders Summit*: Springer, 146-155.

O'Dea, R. E., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W., Parker, T. H., Gurevitch, J., Page, M. J., Stewart, G. and Moher, D. (2021) 'Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: a PRISMA extension', *Biological Reviews,* 96(5), pp. 1695-1722.

Olofsson, H., Brolund, A., Hellberg, C., Silverstein, R., Stenström, K., Österberg, M. and Dagerhamn, J. (2017) 'Can abstract screening workload be reduced using text mining? User experiences of the tool Rayyan', *Research synthesis methods,* 8(3), pp. 275-280.

Onan, A. (2021) 'Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach', *Computer Applications in Engineering Education,* 29(3), pp. 572-589.

Osterwalder, A. and Pigneur, Y. (2010) *Business model generation: a handbook for visionaries, game changers, and challengers.* John Wiley & Sons.

Ouzzani, M., Hammady, H., Fedorowicz, Z. and Elmagarmid, A. (2016) 'Rayyan—a web and mobile app for systematic reviews', *Systematic reviews,* 5(1), pp. 1-10.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A. and Brennan, S. E. (2021) 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', *International journal of surgery*, 88, pp. 105906.

Page, M. J. and Moher, D. (2017) 'Evaluations of the uptake and impact of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement and extensions: a scoping review', *Systematic reviews,* 6(1), pp. 1-14.

Palacios Hidalgo, F. J., Huertas Abril, C. A. and Gómez Parra, M. (2020) 'MOOCs: Origins, concept and didactic applications: A systematic review of the literature (2012–2019)', *Technology, Knowledge and Learning,* 25(4), pp. 853-879.

Paldy, L. G. (2013) 'MOOCs in your future', *Journal of College Science Teaching,* 42(4), pp. 6.

Pappano, L. (2012) 'The Year of the MOOC', *The New York Times,* 2(12), pp. 2012.

Peluso, C. (2022) *Temporal Summarization: a Transformer-Based Approach.* Politecnico di Torino.

Perez-Pena, R. (2012) 'Top universities test the online appeal of free', *The New York Times,* 18, pp. A15.

Pickard, L., Shah, D. and De Simone, J. 'Mapping microcredentials across MOOC platforms'. *2018 Learning With MOOCS (LWMOOCS)*: IEEE, 17-21.

Porter, S. (2015) 'The economics of MOOCs: a sustainable future?', *The Bottom Line,* 28(1/2), pp. 52-62.

Pranckutė, R. (2021) 'Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World', *Publications,* 9(1), pp. 12.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C. D. (2020) 'Stanza: A python natural language processing toolkit for many human languages', *arXiv preprint arXiv:2003.07082*.

Qiao, Y., Xiong, C., Liu, Z. and Liu, Z. (2019) 'Understanding the Behaviors of BERT in Ranking', *arXiv preprint arXiv:1904.07531*.

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q. and Xue, Y. 'Modeling and predicting learning behavior in MOOCs'. *Proceedings of the ninth ACM international conference on web search and data mining*: ACM, 93-102.

Raja, R. and Nagasubramani, P. (2018) 'Impact of modern technology in education', *Journal of Applied and Advanced Research,* 3(1), pp. 33-35.

Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998) *Applied regression analysis: a research tool.* Springer.

Reich, J. and Ruipérez-Valiente, J. A. (2019) 'The MOOC pivot', *Science,* 363(6423), pp. 130-131.

Rish, I. 'An empirical study of the naive Bayes classifier'. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 41-46.

Rizvi, S., Rienties, B., Rogaten, J. and Kizilcec, R. F. (2022) 'Beyond one-size-fits-all in MOOCs: Variation in learning design and persistence of learners in different cultural and socioeconomic contexts', *Computers in Human Behavior,* 126, pp. 106973.

Rohloff, T., Sauer, D. and Meinel, C. 'Students' Achievement of Personalized Learning Objectives in MOOCs'. *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 147-156.

Rõõm, M., Luik, P. and Lepp, M. (2022) 'Learner success and the factors influencing it in computer programming MOOC', *Education and Information Technologies*, pp. 1-19.

Rosasco, L. (2016) 'Introductory Machine Learning Notes', *University of Genoa ML,* 2017.

Rossano, V., Pesare, E. and Roselli, T. (2017) 'Are computer adaptive tests suitable for assessment in MOOCs', *Journal of e-Learning and Knowledge Society,* 13(3).

Ruipérez-Valiente, J. A., Cobos, R., Muñoz-Merino, P. J., Andujar, Á. and Kloos, C. D. 'Early prediction and variable importance of certificate accomplishment in a MOOC'. *European Conference on Massive Open Online Courses*: Springer, 263-272.

Samuelsen, J. and Khalil, M. 'Study effort and student success: a MOOC case study'. *International Conference on Interactive Collaborative Learning*: Springer, 215-226.

Sánchez, M. 'Assessing the quality of MOOC using ISO/IEC 25010'. *2016 XI Latin American Conference on Learning Objects and Technology (LACLO)*: IEEE, 1-4.

Schaffhauser, D. (2018) 'Coursera’s CEO on the evolving meaning of ‘MOOC.’', *Campus Technology*.

Schapire, R. E. (2013) 'Explaining adaboost', *Empirical inference*: Springer, pp. 37-52.

Scpus: Tips and Tricks. Available at: https://blog.scopus.com/tips-and-tricks (Accessed: 24/11/2021 2021).

Sebbaq, H. (2022) 'Fine-tuned BERT Model for Large Scale and Cognitive Classification of MOOCs', *The International Review of Research in Open and Distributed Learning,* 23(2), pp. 170-190.

Shah, D. (2018a) 'The Second Wave of MOOC Hype Is Here, and It’s Online Degrees', Available: Ed Surge. Available at: https://www.edsurge.com/news/2018-05-21-the-second-wave-of-mooc-hype-is-here-and-it-s-online-degrees.

Shah, D. (2018b) 'Six Tiers of MOOC Monetization', *A product at every price point, Class Central. Verfügbar unter https://www. class-central. com/report/six-tiers-mooc-monetization.*

Shah, D. (2019) 'Coursera’s Monetization Journey: From 0 to $100+ Million in Revenue', *ClassCentral. com*.

Shah, D. (2020) 'By The Numbers: MOOCs in 2020'. Available at: https://www.classcentral.com/report/mooc-stats-2020/ (Accessed 15/11/2021).

Shah, D. (2021a) *By The Numbers: MOOCs in 2021*, online. Available at: https://www.classcentral.com/report/mooc-stats-2021/ (Accessed: 02/02/2022).

Shah, D. (2021b) *Massive List of MOOC-based Microcredentials*. Available at: https://www.classcentral.com/report/list-of-mooc-based-microcredentials/ (Accessed: 22/04/2022).

Shah, K., Patel, H., Sanghvi, D. and Shah, M. (2020) 'A comparative analysis of logistic regression, random forest and KNN models for the text classification', *Augmented Human Research*, 5(1), pp. 1-16.

Sharples, M. (2019) 'Visions for the future of educational technology', *EDUCATIONAL VISIONS*, pp. 151.

Shcherbinin, M., Kruchinin, S. V. and Ivanov, A. G. (2019) 'MOOC and MOOC degrees: new learning paradigm and its specifics', *Management Applied. Science Technologies*, 10, pp. 1-14.

Sherstinsky, A. (2020) 'Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network', *Physica D: Nonlinear Phenomena*, 404, pp. 132306.

Shorten, C., Khoshgoftaar, T. M. and Furht, B. (2021) 'Text data augmentation for deep learning', *Journal of big Data*, 8(1), pp. 1-34.

Sievert, C. (2020) *Interactive web-based data visualization with R, plotly, and shiny*. CRC Press.

Simonson, M., Zvacek, S. M. and Smaldino, S. (2019) 'Teaching and learning at a distance: Foundations of distance education 7th edition'.

Singhal, M. S. (2023) 'Predicting Student Performance using Big Data Analysis and Neural Network in Massive Open Online Courses (MOOCs)'.

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y. and Demir, I. (2020) 'A comprehensive review of deep learning applications in hydrology and water resources', *Water Science and Technology*, 82(12), pp. 2635-2670.

Sitanggang, A. B., Putri, J. E., Palupi, N. S., Hatzakis, E., Syamsir, E. and Budijanto, S. (2021) 'Enzymatic Preparation of Bioactive Peptides Exhibiting ACE Inhibitory Activity from Soybean and Velvet Bean: A Systematic Review', *Molecules*, 26(13), pp. 3822.

Song, Y.-Y. and Ying, L. (2015) 'Decision tree methods: applications for classification and prediction', *Shanghai archives of psychiatry*, 27(2), pp. 130.

Springer Link Search Tips. Available at: https://link.springer.com/searchhelp (Accessed: 24/11/2021 2021).

Sraidi, S., Smaili, E. M., Azzouzi, S. and Charaf, M. E. H. 'A sentiment analysis based approach to fight MOOCs’ drop out'. *Networking, Intelligent Systems and Security: Proceedings of NISS 2021*: Springer, 509-520.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014) 'Dropout: a simple way to prevent neural networks from overfitting', *The journal of machine learning research,* 15(1), pp. 1929-1958.

Taneja, S. and Goel, A. (2014) 'MOOC providers and their strategies', *International Journal of Computer Science and Mobile Computing,* 3(5), pp. 222-228.

Teja, P. S. (2019) 'Bagging and Boosting'. Available at: https://medium.com/@saitejaposam9/bagging-and-boosting-2b6cd4a6bda1 (Accessed 06/10/2022).

Tian, Y., Wen, Y., Yi, X., Yang, X. and Miao, Y. 'Predicting Learning Effect by Learner’s Behavior in MOOCs'. *International Conference on Intelligent Data Engineering and Automated Learning*: Springer, 524-533.

Torrey, L. and Shavlik, J. (2010) 'Transfer learning', *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*: IGI global, pp. 242-264.

Tucker, C., Pursel, B. K. and Divinsky, A. 'Mining student-generated textual data in MOOCs and quantifying their effects on student performance and learning outcomes'. *2014 ASEE Annual Conference & Exposition*, 24.907. 1-24.907. 14.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in neural information processing systems,* 30.

Vivian, R., Falkner, K. and Falkner, N. (2014) 'Addressing the challenges of a new digital technologies curriculum: MOOCs as a scalable solution for teacher professional development', *Research in Learning Technology,* 22(1), pp. 24691.

Vrillon, E. (2019) 'Une nouvelle évaluation de la réussite dans les MOOC à partir de registres d’usages individuels', *Questions Vives. Recherches en éducation,* (31).

Wang, L., Hemberg, E. and O’Reilly, U.-M. 'The Influence of Grades on Learning Behavior in MOOCs: Certification vs, Continued Participation'. *2019 IEEE Learning With MOOCS (LWMOOCS)*: IEEE, 122-127.

Wang, L. and Wang, H. 'Learning behavior analysis and dropout rate prediction based on MOOCs data'. *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*: IEEE, 419-423.

Web of Science Core Collection: Search Tips. Available at: https://clarivate.libguides.com/woscc/searchtips (Accessed: 24/11/2021 2021).

Web of Science, Confident research begins here. . Available at: https://clarivate.com/webofsciencegroup/solutions/web-of-science/ (Accessed: 23//11/2021 2021).

Wei, X., Lin, H., Yang, L. and Yu, Y. (2017) 'A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification', *Information,* 8(3), pp. 92.

Wen, M., Yang, D. and Rose, C. 'Sentiment Analysis in MOOC Discussion Forums: What does it tell us?'. *Educational data mining 2014*: Citeseer.

White, S., Davis, H., Dickens, K., León, M. and Sánchez-Vera, M. M. 'MOOCs: What motivates the producers and participants?'. *International Conference on Computer Supported Education*: Springer, 99-114.

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y. and Tingley, D. (2017) 'Delving deeper into MOOC student dropout prediction', *arXiv preprint arXiv:1702.06404*.

Whitehill, J., Williams, J., Lopez, G., Coleman, C. and Reich, J. (2015) 'Beyond prediction: First steps toward automatic intervention in MOOC student stopout', *Available at SSRN 2611750*.

Wintermute, E. H., Cisel, M. and Lindner, A. B. (2021) 'A survival model for course-course interactions in a Massive Open Online Course platform', *PloS one,* 16(1), pp. e0245718.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R. and Funtowicz, M. (2019) 'Huggingface's transformers: State-of-the-art natural language processing', *arXiv preprint arXiv:1910.03771*.

Wong, J., Baars, M., Davis, D., Van Der Zee, T., Houben, G.-J. and Paas, F. (2019) 'Supporting self-regulated learning in online learning environments and MOOCs: A systematic review', *International Journal of Human–Computer Interaction,* 35(4-5), pp. 356-373.

Xiang, F., Zhang, X., Cui, J., Carlin, M. and Song, Y. 'Algorithmic Bias in a Student Success Prediction Models: Two Case Studies'. *2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*: IEEE, 310-315.

Xu, B. and Yang, D. (2016) 'Motivation classification and grade prediction for MOOCs learners', *Computational intelligence and neuroscience,* 2016.

Yang, H. 'Chinese Sentiment Analysis of MOOC Reviews Based on Word Vectors'. *2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)*: IEEE, 68-71.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V. (2019) 'Xlnet: Generalized autoregressive pretraining for language understanding', *Advances in neural information processing systems,* 32.

Yee, M., Roy, A., Stein, J., Perdue, M., Bell, A., Carter, R. and Miyagawa, S. 'The Relationship Between COVID-19 Severity and Computer Science MOOC Learner Achievement: A Preliminary Analysis'. *Proceedings of the Ninth ACM Conference on Learning@ Scale*, 431-435.

Yeomans, M., Reich, J. and Acm (2017) 'Planning Prompts Increase and Forecast Course Completion in Massive Open Online Courses'. *7th International Learning Analytics and Knowledge Conference (LAK)*, Simon Fraser Univ, Vancouver, CANADA, Mar 13-17. NEW YORK: Assoc Computing Machinery, 464-473.

Yin, F., Wang, Y., Liu, J. and Lin, L. (2020) 'The construction of sentiment lexicon based on context-dependent part-of-speech chunks for semantic disambiguation', *IEEE Access*, 8, pp. 63359-63367.

Yin, W., Kann, K., Yu, M. and Schütze, H. (2017) 'Comparative study of CNN and RNN for natural language processing', *arXiv preprint arXiv:1702.01923*.

Yousef, A. M. F., Chatti, M. A., Schroeder, U., Wosnitza, M. and Jakobs, H. 'The state of MOOCs from 2008 to 2014: A critical analysis and future visions'. *International conference on computer supported education*: Springer, 305-327.

Yousef, A. M. F. and Sumner, T. (2021) 'Reflections on the last decade of MOOC research', *Computer Applications in Engineering Education,* 29(4), pp. 648-665.

Zhang, L., Wang, S. and Liu, B. (2018) 'Deep learning for sentiment analysis: A survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 8(4), pp. e1253.

Zhang, Y. and Wallace, B. (2015) 'A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification', *arXiv preprint arXiv:1510.03820*.

Zheng, Q., Chen, L. and Burgos, D. (2018a) 'Certificate Authentication and Credit System of MOOCs in China', *The Development of MOOCs in China*: Springer, pp. 261-276.

Zheng, Q., Chen, L. and Burgos, D. (2018b) 'Emergence and development of MOOCs', *The development of MOOCs in China*: Springer, pp. 11-24.

Zheng, Q., Chen, L. and Burgos, D. (2018c) 'Evaluation Models of MOOCs in China', *The Development of MOOCs in China*: Springer, pp. 207-227.

Zhu, J. and Liu, W. (2020) 'A tale of two databases: The use of Web of Science and Scopus in academic papers', *Scientometrics,* 123(1), pp. 321-335.

Zhu, M. (2021) 'Enhancing MOOC learners’ skills for self-directed learning', *Distance Education,* 42(3), pp. 441-460.

Zhu, M., Sari, A. R. and Lee, M. M. (2020) 'A comprehensive systematic review of MOOC research: Research techniques, topics, and trends from 2009 to 2019', *Educational Technology Research and Development,* 68(4), pp. 1685-1710.